Microarray-based Multiclass Classification using Relative Expression Analysis

by

Sitan Yang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

September, 2014

Abstract

Microarray gene expression profiling has led to a proliferation of statistical learning methods proposed for a variety of problems related to biological and clinical discoveries. One major problem is to identify gene expression-based biological markers for class discovery and prediction of complex diseases such as cancer. For example, expression patterns of genes are discovered to be associated with phenotypes (e.g., classes of disease) through statistical learning models. Early hopes that well-developed methods such as support vector machines would completely revolutionize the field have been moderated by the difficulties of analyzing microarray data. Hence, new and effective approaches need to be developed to address some common limitations encountered by current methods. This thesis is focused on improving statistical learning on microarray data through rank-based methodologies. The relative expression analysis introduced in Chapter 1 is the central concept for methodological development where the relative expression ordering (i.e., the relative ranks of expression levels) of genes is investigated instead of analyzing the actual expression values of individual genes. Supervised learning problems are studied where classification models are built for differentiating disease states. An unsupervised learning task is also examined in which subclasses are discovered by cluster analysis at the molecular level. Both types of problems under study consist of multiple classes.

In Chapter 2, a novel rank-based classifier named Top Scoring Set (TSS) is developed

for microarray classification of multiple disease states. It generalizes the Top Scoring Pair (TSP) method for binary classification problems to the multiclass case. Its main advantage lies in the simplicity and power of its decision rule, which provides transparent boundaries and allows for potential biological interpretations. Since TSS requires a dimension reduction in the training process, a greedy search algorithm is proposed to perform a fast search over the feature space. In addition, ensemble classification based on TSS is also investigated.

In Chapter 3, an efficient and biologically meaningful dimension reduction for the TSS classifier is introduced using the publicly available pathway databases. Pre-defined functional gene groups are analyzed for microarray classification. The pathway-based TSS classifier is validated on an extremely large cohort of leukemia cancer patients. Also, the unsupervised learning ability of relative expression analysis is studied and a rank-based clustering approach is introduced to identify molecularly distinct subtypes of breast cancer patients. Based on the clustering results, the TSP classifier is used for predicting the subtypes of individual breast cancer tumors. These rank-based methods provide an independent validation for the current identification of breast cancer subtypes.

Overall, this thesis provides developments and validations of statistical learning methods based on relative expression analysis.

Primary Reader: Daniel Q. Naiman

Secondary Reader: Donald Geman

Acknowledgments

First it is my pleasure to acknowledge my thesis advisor, Daniel Naiman, for his support and guidance during my graduate study. He introduced me to this interesting topic and constantly encourages me to explore new ideas and methods. His mentorship for me has been outstanding. Second, I would like to acknowledge Donald Geman for his guidance throughout my Ph.D. research, and his insightful comments for my thesis. At the same time, I want to acknowledge my wife, Xi Chen, for her endless support and patience in the past two years. I would not have completed my thesis without her love and encouragement. Also, I would like to acknowledge my parents, Dashan Li and Yizhi Yang, for supporting my study financially and morally. Lastly, my deepest gratitude goes to all of my friends for always making my life enjoyable and pleasant.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Preface

DNA microarray technology was first introduced in the mid-1990s to probe the expression of thousands of genes simultaneously [68, 72]. This technique was quickly adopted for the study of a wide range of biological processes. The use of gene expression microarray data initially focused on the identification of differentially expressed genes associated with human diseases (e.g., asthma, heart disease and cancer) [14, 27]. In the last decade, the application of gene expression microarrays to discover and discriminate subclasses of human cancer has attracted tremendous interest. Cancers are conventionally classified by the type of tissue in which the cancer originates. However, the traditional histopathology of a cancer specimen by experienced doctors is subjective and prone to human error. Moreover, different cancer subclasses can have similar histopathological appearances but may differ substantially in terms of therapeutic responses and clinical courses [35]. In this situation, microarray-based gene expression profiling offers a hope that statistical analysis tools can be developed to improve the diagnosis and prognosis of cancer.

Early studies used unsupervised methods such as cluster analysis to identify subgroups of patients based on gene expression patterns [2]. Such studies suffered from the limitation that only samples in large retrospective studies could be classified by unsupervised methods, and results provided little guidance for clinical prognosis of individual samples. In 1999, Golub et al. [35] pioneered a supervised molecular classification of cancer using gene expression profiling. They devised a statistical scheme based on patient samples having two subtypes of leukemia cancer: acute lymphoblastic leukemia and acute myeloblastic leukemia. The diagnostic type of each patient was known a priori. They demonstrated the ability of the scheme to assign appropriate subtype labels for another independent group of leukemia patient samples. This study received wide attention in the research community and inspired subsequent studies to search for predictive gene expression patterns using statistical tools.

As a result, numerous statistical methods were introduced for microarray-based disease classification. Most of these methods are drawn from other research areas that require data analysis such as computational vision and natural language processing. In general, they are so called "machine learning" methods that aim to make a certain kind of prediction on the new data. In the context of disease classification, a simple example would be to predict whether or not a particular patient has a certain disease. While such binary classification problems were widely explored in the scientific literature and achieved considerable success [55, 92], problems that contain multiple classes received relatively less attention [31, 79]. However, as the genetic heterogeneity of complex diseases such as cancer has been increasingly appreciated in recent studies [8, 75], the discovery of new disease classes is expected to continue, leading to a growing number of multiclass problems. Moreover, a large number of experiments for investigating diseases in

stage, survival time and therapeutic response are producing microarray data encompassing multiple classes [9,21]. As a result, there is an increasing need for developing statistical methods that can handle multiclass problems.

This thesis is focused on addressing the limitations of current machine learning techniques as applied to disease classification of multiclass problems using microarray analysis. In particular, common issues such as the small-sample learning situation and the interpretability of classification rules are discussed in detail. In this chapter, the biological and statistical background of problems investigated in this thesis are first described. Then, the motivation and fundamentals to develop new statistical methods for multiclass problems are presented. Next, an overview of proposed methods and contributions is provided.

1.2 Microarray Data

A microarray consists of a dense patterning of thousands of cDNA strands or oligonucleotide sequences immobilized on an inert substrate (usually a glass slide or a nylon membrane). The immobilized sequences are referred to as "probes". The microarray technique uses gene-specific probes to detect variations in levels of gene expression in a target biological sample. As described in [68], Gene expression experiments start by isolating RNA from tissue of interest and tagging it with a detectable marker (e.g. fluorescent dye). The labeled RNA is then washed over the surface of the microarray, and allowed to incubate for a period of time, during which samples of messenger RNA (mRNA) hybridize to the target gene probes. Then scanning of the microarray, using either confocal microscopy or phosphor imaging techniques, yields quantifiable digital images of the array hybridization results. The relative fluorescence intensity of each gene-specific probe

is a measure of the level of expression of the particular gene. The greater the degree of hybridization implies a higher relative intensity.

The data collected after scanning consists of intensity readings for each spot on the array. These intensity readings are typically biased and noisy measurements of gene expression levels due to systematic and random variations in the microarray experiment, and they often need to be carefully preprocessed, with different microarray platforms requiring different preprocessing methods. These methods share three common steps: background adjustment, normalization and filtering. Background adjustment corrects for non-specific hybridization and noise in the optical detection system. Normalization aims to facilitate the comparison between different arrays. It compensates for variations such as efficiencies of labeling, hybridization reactions, array differences and laboratory conditions. Additionally, the data are often filtered to exclude genes that show limited variation or relatively low intensity in expression levels across all biologic samples. In general, there are a variety of statistical analysis methods proposed for preprocessing micrarrays in the literature [44, 68].

After preprocessing, the data are usually transformed into a matrix of expression values. Each row of the matrix contains the expression values of a gene and each column represents a specific biological sample. The expression values are positively correlated with the intensity readings so that a relative large value indicates the gene is highly expressed compared to other genes. However, there are no apparent patterns that can be immediately identified to associate with any group of samples. In fact, the expression matrix often contains tens of thousands of genes but only up to hundreds of samples. Most genes seem to provide noisy and inconsistent signals across all samples, and it is extremely difficult to study all genes as a whole given the gene interdependencies and

the small sample size available. Therefore, a routine analysis is to find genes that show patterns of expression correlated with disease states. Such groups of genes are typically referred to as "gene signatures" where their diagnostic and prognostic value are assessed for a variety of clinical applications.

1.3 Statistical Background

Biological insights for various diseases can be gained by exploring microarray data. In particular, many problems of studying diseases appear in the form of "supervised learning". In a general supervised learning problem, there is often a set of variables that are measured as inputs and they will have some influence on one or more outputs. Formally, the inputs are often called features, and depending on the type of the problem, the outputs are referred to as class variables in classification or responses in regression analysis. For either case, the purpose is to use the features to predict the values of the outputs. In the context of supervised classification, the features are the expression levels for the set of genes from microarray data, and the class variable takes categorical values representing different patient groups identified based on prior knowledge or different outcomes of clinical trials. Since the purpose of classification is always to establish a rule that can be used for prediction, these two terms ("classification" and "prediction") are used interchangeably throughout the thesis. Here, the classic Bayesian decision theory is first presented to provide a framework for developing statistical models used in supervised classification. Then an important concept is discussed related to the assessment of classification methods known as "Bias-Variance Tradeoff".

1.3.1 Decision Theory

Let $X \in \mathbb{R}^p$ denote an ordered set of real valued random variables, and Y a discrete random variable. In particular, X is a $p-$dimensional random vector consisting of features and Y is the corresponding class variable. The goal is to seek a function $f(X)$ to predict Y given the values of variables in X. Typically $L = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ represents a set of N training samples observed before hand on which a classifier/predictor $\hat{f}(X)$ is trained. Therefore, each observed x of X makes a prediction of its class label as $\hat{f}(x)$. In theory, a loss function $l(\hat{f}(X), Y)$ is required here for penalizing misclassification errors. A commonly used loss function in classification is the 0-1 loss

$$l(\hat{f}(x), y) = \mathbb{I}_{\{\hat{f}(x) \neq y\}},$$

which assigns one unit of loss for each misclassification. After specifying a loss function, an key quantity of interest is the error incurred in prediction, also referred to as the "generalization error", which is the expected loss over a set of samples independent of those used for training

$$E[l(\hat{f}(X), Y)],$$

where both X and Y are drawn from their joint probability distribution $P(X, Y)$. In this situation, the best classifier $\tilde{f}$ is the one that minimizes the generalization error

$$\tilde{f} = \arg\min_{\hat{f}} \ \mathrm{E}[l(\hat{f}(X), Y)].$$

To obtain $\tilde{f}$, it suffices to consider the minimum of the error at each point x

$$\tilde{f} = \arg\min_{\hat{f}} \ \mathrm{E}[l(\hat{f}(X), Y)|X = x].$$

Now suppose $Y \in \{1, 2, ..., K\}$ and $\hat{y} = \hat{f}(x)$, the equation above can be written as

$$\begin{aligned}
\tilde{f} &= \arg\min_{\hat{f}} \sum_{k=1}^{K} l(\hat{f}(x), k) \cdot P(k|X = x) \\
&= \arg\min_{\hat{y} \in \{1,2,...,K\}} \sum_{k=1}^{K} \mathbb{I}_{\{\hat{y} \neq k\}} \cdot P(k|X = x) \\
&= \arg\min_{\hat{y} \in \{1,2,...,K\}} 1 - P(\hat{y}|X = x) \\
&= \arg\max_{\hat{y} \in \{1,2,...,K\}} P(\hat{y}|X = x).
\end{aligned}$$

This optimal solution is often known as the "Bayes rule", and it concludes that the best prediction for Y is the class that maximizes the posterior distribution given X. This rule is also commonly referred to as *maximum a posteriori*. In practice, this posterior distribution is often unknown and has to be estimated from the data. According to Bayes' theorem,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)},$$

and so

$$\arg\max_{y} P(y|X) = \arg\max_{y} P(X|y)P(y).$$

Thus, an alternative solution is to estimate $P(X|Y)$ and $P(Y)$, the class conditional densities and the prior distribution of the class variable.

The decision theory discussed above serves as the foundation from which many classification methods are derived for modeling either $P(X|Y)$ ($P(Y)$ often estimated as the observed frequencies or simply equal priors) or $P(Y|X)$. For the former case, a classic example is the linear discriminant analysis classifier that assumes each class density $P(X|Y = k)$ is a p-dimensional multivariate Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$ where the Σ_k takes a common value Σ, for all k. Another example is the naive Bayes classifier, which makes a simplified assumption that each of the conditional class densities are products of marginal densities of features. For the Gaussian naive Bayes classifier, each marginal

density is assumed to be Gaussian, and in this case $P(X|Y = k)$ turns out to be a product of many one dimensional Gaussian densities. On the other hand, there exist a number of methods that aim to model $P(Y|X)$ directly, and the k-nearest-neighbor classifier is one of them. In this method, the posterior probabilities at one point $P(Y|X = x)$ are actually relaxed to posterior probabilities within a neighborhood of a point $P(Y|X \in N_k(x))$, which are empirically estimated by training-sample proportions. The logistic regression classifier is another example. It arises from the motivation to model the log-odds ratios of posterior probabilities as linear functions of x given by

$$\log \frac{P(Y = k|X = x)}{P(Y = K|X = x)} = \alpha_k + \beta_k^T x \quad k = 1, 2, ..., K - 1,$$

where $\alpha_k \in \mathbb{R}$ and $\beta_k \in \mathbb{R}^p$ are model parameters. Then a simple calculation shows that

$$P(Y = k|X = x) = \frac{\exp(\alpha_k + \beta_k^T x)}{1 + \sum_{i=1}^{K-1} \exp(\alpha_i + \beta_i^T x)}, \ k = 1, 2, ..., K - 1,$$
$$P(Y = K|X = x) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\alpha_i + \beta_i^T x)}.$$

Therefore, these posterior probabilities are immediately available given the estimates of model parameters.

In general, after approximating $P(X|Y)$ or $P(Y|X)$ from the data, the Bayes rule can be constructed for classification. However, this empirical rule may no longer be optimal due to the assumptions made for modeling either $P(X|Y)$ or $P(Y|X)$. Thus, the quality of a particular classification model can depend on how well the model assumptions fit the data. However, even if the model is correct, the calibration of the model can still be limited by the sample size available. Regardless of the sample size issue, the misspecified model assumptions can pose severe limitations on many classification models because the microarray data typically follow complex and unknown distributions that are likely to violate the assumptions of these models. In this situation, classifiers that do not rely on

the theory described above may produce an improved performance. In fact, there exists a collection of methods called "discriminative" methods that do not rely on estimates for either $P(X|Y)$ or $P(Y|X)$. Instead, they directly focus on identifying effective decision boundaries of different classes. Probably the most popular discriminative method is support vector machine (SVM) [11]. SVM originates from the perceptron classifier [74] developed in the late 1950s that constructs a hyperplane via a linear combination of the input features to separate two different classes, and the hyperplane is found by minimizing some target function related to misclassification. In contrast, SVM usually produces nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space. Therefore, SVM provides a large degree of flexibility and can handle difficult classification problems.

1.3.2 Bias-Variance Tradeoff

Each of the classification approaches described above has a certain level of complexity. Different approaches may have different levels of complexity because they are based on different underlying theories and they differ in the type and number of parameters. The complexity of an approach generally affects its predictive performance. For example, nonlinear classifiers may be more powerful than linear classifiers for a classification problem that has complex decision boundaries. In this sense, it is plausible that a model's performance can be improved by increasing complexity. However, as explained by the tradeoff between bias and variance described below, this is not always the case.

As in the previous section, the goal is to find a function $f(X)$ to predict Y. The focus is to analyze the generalization error incurred, which can depend on the choice of loss

functions. The generalization error of prediction has the form of

$$E[l(\hat{f}(X), Y)]$$

where $\hat{f}$ is the classification model estimated from the training data. Note that this expectation is averaged over the training data that produce $\hat{f}$. Assume $\hat{Y} = \hat{f}(X)$ and $Y \in \{1, 2, ..., K\}$, the bias-variance decomposition of $E[l(\hat{Y}, Y)]$ for the quadratic loss function is actually well known, e.g., see [42] and serves as an important tool for analyzing learning algorithms. For classification problems where the 0-1 loss function is a common choice, the bias-variance decomposition of $E[l(\hat{Y}, Y)]$ has also been carefully investigated, e.g., see Dietterich and Kong [19], Breiman [6] and Kohavi and Wolpert [51]. Regardless of the detailed mathematical descriptions in these studies, the generalization error typically has three components: σ^2, Bias^2 and Variance. The σ^2 term is sometimes called "irreducible error" because it is inherently caused by the target Y and is independent of the learning algorithm. Thus, the quantities of interest are often the "Bias^2" and the "Variance" term. For the quadratic loss function, the "Bias^2" term represents the error incurred due to the limited capacity of the model family to approximate the true relation between Y and X, and the "Variance" term explains the error caused by estimating the model from the training data. As mentioned above, these terms are also shared by the 0-1 loss function, but may have different explanations. According to [51], the "Bias^2" term measures the squared difference between the target Y and the algorithm's output $\hat{Y}$ averaged over all possible training sets, and the "Variance" term measures how sensitive the algorithm is to the changes in the training set. The variance reaches the minimum value zero if the learning algorithm always makes the same guess regardless of the training set. Regardless of the type of the loss function, the model complexity links the bias and the variance terms by introducing a tradeoff between them. As the model

complexity is increased, the bias tends to decrease, but typically the variance increases. For example, a complex model may have a complex decision boundary that adapts better than a simple model to the probability change of the target Y for different sample points, but the complexity of the model may result in sensitivity to the change in the training data and cause an increase of the variance, which leads to so called "over-fitting" problem.

For problems considered in this thesis, "small n, large p" is a common situation. Here n refers to the number of training samples and p represents the number of features, or specifically, genes. In this case, supervised learning is a formidable task because the high-dimensional data requires complex models while the small sample size demands simple models. This "small n, large p" dilemma widely exists in problems of computational biology where many traditional statistical methods derived from other fields often break down. In particular, these methods are generally designed for producing accurate predictions under the assumption that there are "enough" training samples, a condition that holds in many other fields such as image analysis or natural language processing. However, learning from biological data based on these methods is likely to cause over-fitting. The small training-sample size barely provides good estimates of the model parameters, which can be subject to large variances. Hence, in view of the bias-variance tradeoff, the variance term here plays a much more important role, and it is necessary to restrict the model complexity. A possible strategy is to develop simple models based on plausible biological assumptions so that the model bias can be compensated by the variance reduction.

1.4 Relative Expression Analysis

The main challenge for microarray-based classification is to develop techniques that yield relatively accurate and robust decision rules. Since gene expression measurements largely depend on microarray experiments, microarray-based statistical inference can be affected by a number of factors, such as experimental design and the type of data normalization. Therefore, traditional techniques in statistical analysis may need to be redesigned to address these issues. Also, there is always value added if a particular decision rule can be interpreted with respect to the biology of the underlying disease. Interpretability facilitates follow-up studies for biological validation, which may eventually lead to the development of clinical applications. However, many advanced classification techniques such as support vector machines [11], random forest [7] and neural networks [49] are derived for general purpose classification where accuracy is the top priority. Such models can be viewed as "black boxes" where the decision process is hidden and can be quite complex, and their decision rules do not lend easily to biological mechanistic understanding.

For complex diseases, the phenomenon of interactions among multiple genes and other molecular agents within biological networks can be expected to be considerable. Genes in networks connect and operate in a combinatorial manner. Hence the information gleaned from the expression patterns of individual genes may only serve as references for each other and become biologically meaningful only in a relative comparison. Recently, methods for analyzing microarray data based on biologically meaningful pathways or networks have achieved considerable success [32, 50, 84]. These frameworks have been applied to diverse cancer systems and serve as a robust source of biological discovery.

One way to probe the interactions among genes is to study the relative ordering

(ranks) of their expression values. In microarray experiment, expression values within a sample are assayed simultaneously, and the relative ordering of expression can be more reliable and robust than raw values: it is likely to be preserved under slight perturbations of gene expression values and are robust against any effect that shifts expression values in the same direction. Lin [90] showed that the relative ordering is invariant under commonly used preprocessing techniques such as convolution and quantile normalization in RMA [44].

The idea of using relative ordering was proposed by Geman et al. [33] in which the expression values of two genes were compared for distinguishing two phenotypes of interest. A rank-based classification approach, named "Top Scoring Pair"(TSP), provided transparent but powerful decision rules which were shown to compete with many sophisticated machine learning approaches. In subsequent studies, TSP has been extended in a number of ways. Xu et al. [89] considered the average ranks in two groups of genes (rather than a pair of genes) for constructing the decision rule. Tan et al. [83] introduced the k-TSP classifier where the top k scoring pairs (i.e., gene pairs with top k scores) were involved using the majority rule in the decision process. Also, Lin et al. [56] proposed the "Top Scoring Triplet" method in which relative ordering in each triplet (i.e., three genes) were investigated using a similar approach in TSP. Recently, Kaur et al. [47] introduced the "ProtPair" method that used TSP for human disease prognosis based on protein expression data. Such rank-based methods are described as forms of what is termed "Relative Expression Analysis" in [23].

In principal, relative expression analysis investigates the combinatorial patterns associated with a group of genes. Probability distributions of different combinatorial expression orderings are often estimated for small-sized gene sets and statistical significance is

assessed based on estimates. The ordering of expression values is obtained within each sample and no cross-sample comparisons are considered. Although such a comparison may not represent the actual mechanisms of complex diseases, it is likely to be linked with biological activities such as regulation of gene expression. In fact, given the amount of data typically available for microarray analysis, the empirical distribution of the ordering of expression values for a small collection of genes seems to be one of the few types of statistics that can be robustly estimated. Overall, relative expression analysis methods provide simplified models to study genetic networks and can potentially offer insights into biological mechanisms.

1.5 Proposed Methodology

This thesis focuses on developing rank-based methods for statistical learning problems arising in gene expression analysis. In particular, methods are derived to improve microarray-based classification in the presence of multiple classes. The discriminative power of small-sized gene sets among phenotypes are explored where the relative expression in each gene set is the focus rather than the actual values.

The first attempt is to extend the TSP method in the multiclass setting. Previous extensions discussed in the last section are all focused on binary problems. In the multiclass case, the "small n, large p" problem is especially compounded when subdivision of an already small set of samples into subclasses leads to dramatically smaller sample sizes for subclasses. Also, multiclass methods typically require significantly more computation, and decision rules generated can become more complex. To address these issues, the "Top Scoring Set" method is proposed to generalize the idea of relative comparison in TSP. For an m-class problem, TSS searches for a number of m-gene sets to make class predictions.

Each selected set (i.e., top scoring set) tends to preserve a distinct ordering of its member genes in a single class, which is considered as the signal for class discrimination.

The second attempt focuses on discovering similarity of samples based on relative ordering of their expression values. In particular, a rank-based clustering approach is considered that uses the Kendall's rank coefficient [48] as the distance measure for any pair of samples. After the distance matrix is calculated, the traditional hierarchical cluster analysis is applied. This approach aims to quantify the differences among classes by investigating combinatorial behaviors of gene pairs, and can also be combined with supervised classification methods.

At the same time, attempts have also been made to develop techniques to improve the efficiency of training classification methods proposed in the thesis. Although conventional feature selection methods such as one-way ANOVA or Kruskal-Wallis, its non-parametric analog can be used, it is better to have techniques particularly suited to microarray data and problems under study. In this thesis, two such approaches have been developed. The first one applies a greedy search algorithm on the expression values of individual genes. The second one tries to incorporate biological knowledge when pathway information can be inferred from the gene expression data.

1.6 Summary of Contributions

In summary, this thesis brings forth the following research developments:

1. Development of a new approach for microarray-based classification of multiclass problems mainly related to cancer. The approach naturally handles multiclass microarray data and acquires some nice properties such as parameter-free-ness and

invariance under several common preprocessing techniques. The decision rule is transparent and easy to interpret. This classification approach has been validated on seven microarray gene expression data of human cancers including leukemia, lung and bladder cancer. Its robustness has been demonstrated on an extremely large cohort of leukemia cancer patients and a cross-center cohort of bladder cancer patients. In addition, the potential of the approach to combine ensemble techniques such as boosting has also been explored.

2. Proposal of two methods to improve the efficiency of training the classification approaches introduced above. The first method uses a greedy search algorithm to quickly screen the entire set of genes to search for discriminative expression patterns. The second one relies on pre-defined gene groups from public databases such as pathways to restrict the search. These two methods are tested on a variety of microarray data sets.

3. Proposal of a rank-based clustering method used for unsupervised learning tasks with respect to microarray analysis. The method explores a group of pre-selected genes that are potentially related to phenotypes of interest, and constructs a distance matrix for samples based on the Kendall's rank coefficient. Hierarchical cluster analysis is then built upon the distance matrix. This method is validated using a compendium of three breast cancer data sets where it is applied to identify distinct subtypes for analyzing patient survival time. The prognostic value of these identified subtypes is compared to published expression-based models and gene signatures.

In general, this thesis focuses on a variety of multiclass problems with respect to gene expression analysis. It demonstrates the importance of tailoring statistical learning

methods to microarray classification and the usefulness of developing methodologies based on relative expression analysis.

Chapter 2

Top Scoring Set

2.1 Introduction

The study of complex diseases such as cancer via microarray analysis is producing an ever-growing set of multiclass classification problems. As limitations (e.g., reproducibility and interpretability) of traditional machine learning techniques have been increasingly appreciated, new and effective methodologies are expected to be developed to address these problems. In this chapter, the "Top Scoring Set" (TSS) approach [91] is developed for classification of microarray data containing multiple classes. The background of multiclass classification in microarray analysis is first introduced. Then the TSS classifier is described and discussed in detail, followed by two validation studies. Lastly, some theoretical developments of TSS are presented.

2.1.1 Multiclass Methods

The literature on microarray-based classification methods is extensive, see, e.g., the review by Pirooznia et al. [66]. Some discussions of multiclass classification methods for

gene expression analysis have been provided by Statnikov et al. [79] and Tao et al. [55]. These multiclass approaches can be roughly divided into direct and indirect approaches. Direct approaches can immediately apply to multiclass problems, while indirect ones rely on schemes such as "one-vs-one" [30] or "one-vs-all" [74] to decompose a particular multiclass problem into a set of binary problems. Thus, indirect approaches tend to be more computationally intensive because they aim to solve an ensemble of simpler problems.

The most straightforward direct approach is k-Nearest-Neighbor (kNN). The main idea of kNN is based on the concept of similarity. The distance between the test sample whose class is to be decided and each of the training samples is computed using a certain metric (e.g., Euclidean distance). To predict a test sample, kNN uses the class labels of the k closest samples in distance from the training set and takes a majority vote. Therefore, kNN can naturally handle multiple classes and hence often serves as a benchmark method in many studies [55,83,92]. It also has been proven that the error of kNN is asymptotically at most two times the optimal Bayesian error [18].

Another commonly used direct approach is Naive Bayes (NB). NB assumes that given any class, the distributions of features are independent from each other. As a result, it expresses the multivariate conditional distribution of features given the class as a product of the conditional distributions of the individual features. Hence, NB only requires estimation of a number of parameters that is linear in the dimension of feature space. Although this assumption is often violated in practice, it has at least two advantages. First, it is quite computationally efficient for high-dimensional microarray data, and second, it reduces variance error by possibly creating bias error, though the bias error may be arbitrarily large. The "naive" assumption also allows for explicit derivation of the

Bayes rule for class prediction, and this rule can be applied immediately to any number of classes.

Linear discriminant analysis (LDA) and its variants constitute another classic family of direct methods. Typically they assume the class conditional distributions as multivariate Gaussian where class-wise mean values differ but class-wise covariance matrices are the same. For microarray classification, LDA is greatly challenged by the "small n, large p" problem due to the number of correlations among features to be determined versus the size of available samples. Therefore in practice, either a small set of features has to be selected prior to classifier training or some variants of LDA such as diagonal LDA are considered. A good improvement of LDA as applied to gene expression analysis is proposed by Tibshirani et al. [85]. The method is known as "Prediction Analysis of Microarray" (PAM). It eliminates the effect of many noisy genes in traditional discriminant analysis by shrinking their values towards the class centroids, i.e., the mean expression values of classes. As a result, PAM uses relatively few genes for classification, which is quite favorable for understanding the underlying biology.

Decision trees (DT) are also commonly used for multiclass problems. An early implementation of decision trees, known as CART (Classification and Regression Tree), used binary decisions. A popular later development is C4.5 tree by Quinlan [69]. DT sequentially selects important features as split points and reaches its decision as adequate splits are made. The choice of split points is often based on concepts of information theory such as entropy. For applying DT on microarray data, one main advantage is that it offers an interpretable decision rule that consists of "if...else..." rules when tracing the tree from top to bottom. The normal decision tree can grow to a large size and it is often pruned to control the computational cost and to address over-fitting. But the pruned

tree only utilizes a small set of features and the model bias can be large. This limitation is addressed by "random forests" (RF) [7], an ensemble classification method that operates by constructing multiple decision trees based on random choices of features on the training data. The final class prediction is the mode of the classes output by individual trees. The power of RF has been demonstrated in many studies [53, 80].

Support vector machines (SVMs) [11] are perhaps the single most important development in supervised classification. They have been proved to obtain superior classification performance compared to other methods in many domains and tasks, and they can handle large-scale classification in both samples and variables. In particular for gene expression analysis, their powerful performance has been demonstrated in various studies [62, 80]. SVMs originate from optimal hyperplane classifiers such as perceptrons [70]. Samples in the original space are projected onto a higher dimensional feature space where they can be well separated by a maximal margin hyperplane (the margin is the distance between the hyperplane and the sample closest to it), which can be found by solving an optimization problem. SVMs were initially designed for binary classification problems, and there are two major approaches to extend them in the multiclass setting. The first approach is to use decomposition schemes mentioned above to reduce the problem into a set of binary sub-classification problems, which are solved individually and their decisions lead to the final decision based on a majority vote. The second approach is to modify the original (binary) optimization equation into a multiclass objective function [12, 43]. These methods are often referred to as multiclass SVMs.

Unlike direct approaches, indirect approaches require the decomposition of a multiclass problem into binary ones. "One-vs-all" is a popular scheme that decomposes a K-class problem into a set of K sub-problems, each of which is a binary problem. To be

precise, for class $i = 1, 2, ..., K$, a classifier is constructed to distinguish between i and $\{1, 2, ..., i-1, i+1, ..., K\}$. The final decision for class prediction is the class that receives the most votes from these K classifiers. As discussed before, another commonly used scheme is "one-vs-one" and is also known as pairwise coupling. In this case, for every distinct pair of classes i and j, $i, j \in \{1, 2, ...K\}, i \neq j$, a binary classifier is trained using samples from those two classes. As a result, $\frac{K(K-1)}{2}$ classifiers are constructed, each of which votes for a single class. The final decision again is the class that gets the most votes.

In summary, this section provides an extensive list that covers most of the current multiclass methods, many of which serve as benchmark methods in this thesis.

2.1.2 Related Work

Supervised classification problems based on microarray analysis have been investigated for years. Standard methods in areas such as machine learning and pattern recognition are routinely applied to microarray data, including neural networks [49], decision trees and support vector machines [80]. But the learning abilities of these state-of-the-art techniques are often limited due to the "small n, large p" dilemma. In view of the bias and variance trade-off, simplifying assumptions and special designs appear necessary.

One of the approaches attempting to address these limitations was proposed by Geman et al. [33]. They introduced the "Top Scoring Pair" (TSP) method as a new binary classification approach by simply comparing expression levels in one or more pairs of genes (i.e., top scoring pairs) to make a classification decision. As illustrated in Figure 2.1, the expression levels of gene *SPTAN1* and *CD33* are displayed for 72 patient samples from [35], which are grouped according to two types of leukemia cancer: acute myeloid leukemia

(AML) and acute lymphoblastic leukemia (ALL), with 25 and 47 samples respectively. TSP classifies a sample as ALL if *SPTAN1* has higher expression level than that of *CD33* in the sample, and AML otherwise. Although this decision rule seems simple, it was demonstrated to compete with a number of sophisticated machine learning approaches. In addition, gene pairs selected by TSP in various subsequent studies were found to be biologically informative, see, e.g., [24], [94] and [63].

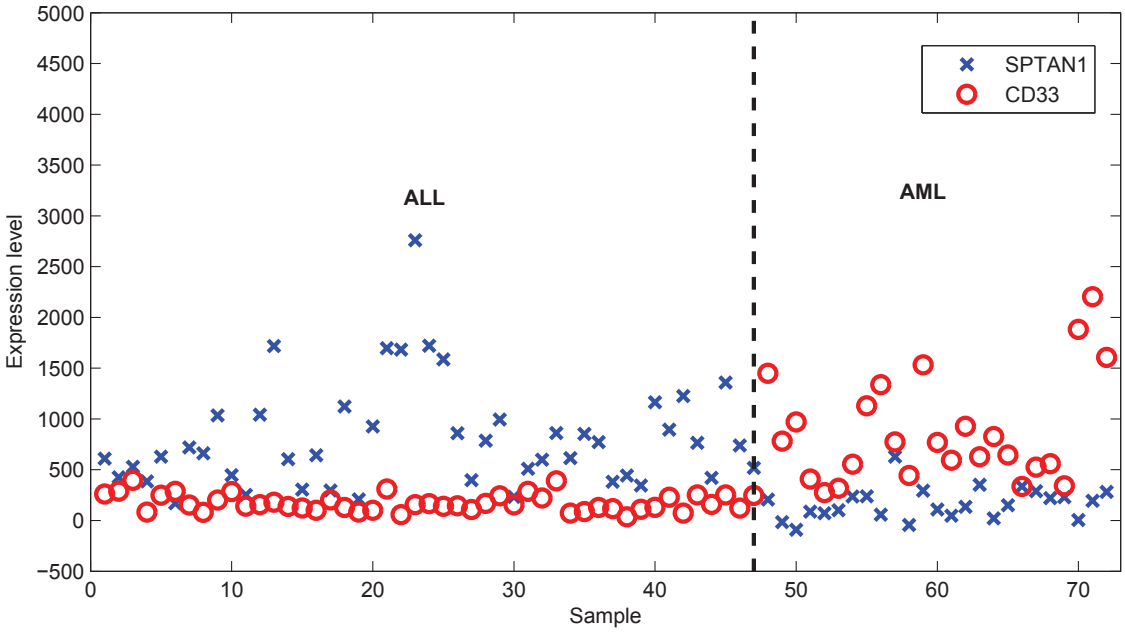

Figure 2.1: Gene expression patterns for a top scoring pair of genes.

The TSS approach developed in this chapter is motivated by TSP and is essentially a generalization of TSP in the multiclass case. For an m-class problem, the class prediction is determined by a relatively small number of m-gene sets, namely, top scoring sets. Each top scoring set votes for a class based on the ordering of expression levels of its genes. The final prediction is the class that receives the majority of votes. In principle, TSS makes specific statistical hypotheses about gene expression comparisons that could have biologic interpretations, and even without the potential interpretability, the decision rule itself can be easily appreciated by non-specialists.

There is no question that some gene expression patterns that are potentially useful for classification may be dismissed by TSS, and the assumption made in TSS might seem insufficient to reflect biological conditions in complex diseases. However, TSS provides a practical attempt for modeling the statistical dependency structure among genes given the amount of data, and results achieved in this chapter demonstrate that the information in the ordering of gene expressions is sufficient to reliably perform classification.

2.2 Methods

Consider G genes measured using DNA microarrays and their expression levels $X = \{X_1, X_2, ..., X_G\}$ regarded as a random vector. Each observed gene profile x is a realization of X and has a true label y representing its class. A microarray data set is a collection of many, say N, observed gene profiles and can be represented as a matrix $\{x_{ij}\}$ with G rows of genes and N columns of samples (typically $G \gg N$). This section begins with a brief introduction of the TSP method. Then the "Top Scoring Set" classifier is developed as a new approach by generalizing TSP in the multiclass setting.

2.2.1 A short review of TSP

As discussed in [33], for a two-class problem (with classes denoted by 1 and 2), TSP aims to find each "marker" gene pair (i, j) $(i, j \in \{1, 2, ..., G\})$ that has a simple relation whose probability distribution changes significantly from one class to the other. The simple relation considered here is the comparison between the expression levels of gene i and j, and a highly relevant quantity of interest is $P(X_i > X_j \mid y)$ where y is the class variable, $y \in \{1, 2\}$. So if $P(X_i > X_j \mid y = 1)$ is high while $P(X_i > X_j \mid y = 2)$ is low, it

will be very likely to observe $X_i > X_j$ in class 1 but not in class 2 where $X_i < X_j$ is more likely to happen. As a result, this property of (i, j) leads to the ability to distinguish between two classes simply by determining the gene having the higher expression value, a simple decision rule for predicting class labels. In TSP, a score is defined for each distinct gene pair (i, j) as $|\hat{P}(X_i > X_j \mid y = 1) - \hat{P}(X_i > X_j \mid y = 2)|$ in order to estimate the probability change from class to class where $\hat{P}(X_i > X_j \mid y)$ is the frequency observed from the data. Those that achieve the highest score among all possible gene pairs (i.e., top scoring pairs) are involved in the decision rule. For a top scoring pair that has $\hat{P}(X_i > X_j \mid y = 1) > \hat{P}(X_i > X_j \mid y = 2)$, it predicts the class label $\hat{y}$ of a new sample x as

$$\hat{y} = \begin{cases} \text{class 1, if } x_i > x_j, \\ \\ \text{class 2, if } x_i < x_j. \end{cases} \tag{2.2.1}$$

Then the predictions for each class are summed up and the majority rule is applied to produce the final prediction. The decision rule of TSP is only based on simple comparisons of gene pairs. However, it has been shown as an effective classifier on many cancer data sets, and some top gene pairs from these studies have been found to be informative. Also, TSP has been extended in a number of ways. Xu et al. [89] considered the average ranks in two groups of genes (rather than a pair of genes) for constructing the decision rule. Tan et al. [83] introduced the k-TSP classifier where the top k scoring pairs are involved using the majority rule in the decision process. Also, Lin et al. [57] proposed the "Top Scoring Triplet" method in which relative orderings in each triplet (i.e., three genes) are investigated using a similar approach in TSP. Recently, Kaur et al. [47] introduced the "ProtPair" method that uses TSP for human disease prognosis based on protein expression data. All of these derivations have been for binary problems so far.

2.2.2 Top Scoring Set

This section introduces TSS as a new multiclass classification approach. The motivation of TSS comes from the relative comparison idea used in TSP. As discussed in [33], relative comparison of mRNA concentrations indicated by gene expression levels provides a natural link with biochemical activity, and proposes concrete hypotheses for a small list of genes. Therefore, the goal here is to discover valuable information for separating multiple classes by comparing expression patterns of a few genes. In particular, for an m-class problem (with classes denoted by 1,2,...,m), m "marker" genes $\mathcal{S} = \{i_1, i_2, ..., i_m\} \subset \{1, 2..., G\}$ are found in which the presence of some simple relations among these genes with high conditional probability depending on the class leads to class separability. Specifically, the high expression level of gene i_c relative to the other $m-1$ genes in $\mathcal{S}$ is assumed to be strongly indicative of a sample coming from class c. To be precise, the desired statistical property for $\mathcal{S}$ is that $\forall c \in \{1, 2, ..., m\}$

$$P[\arg\max\{X_r, r \in \mathcal{S}\} = i_c \mid y = c] \gg P[\arg\max\{X_r, r \in \mathcal{S}\} = i_c \mid y \neq c]. \qquad (2.2.2)$$

In other words, gene i_c is much more likely to be the gene that has the maximum expression level in $\mathcal{S}$ for class c than for any other class. In this case, a classification rule can be constructed by determining which gene is most expressed in $\mathcal{S}$ with a simple "max" function. Therefore, it is essential to find gene sets satisfying (2.2.2) with the greatest possibility for discrimination. For this purpose, a score is defined for each m-gene set to estimate its probability of holding (2.2.2). The sets with the highest score are hence referred to as the top scoring sets and are used for classification.

An example of a TSS classifier is illustrated in Figure 2.2. Here a more difficult task than that in Figure 2.1 is considered to distinguish three subtypes of the leukemia data from [35]. Three leukemia subtypes are AML, B-ALL (B-cell ALL) and T-ALL (T-cell

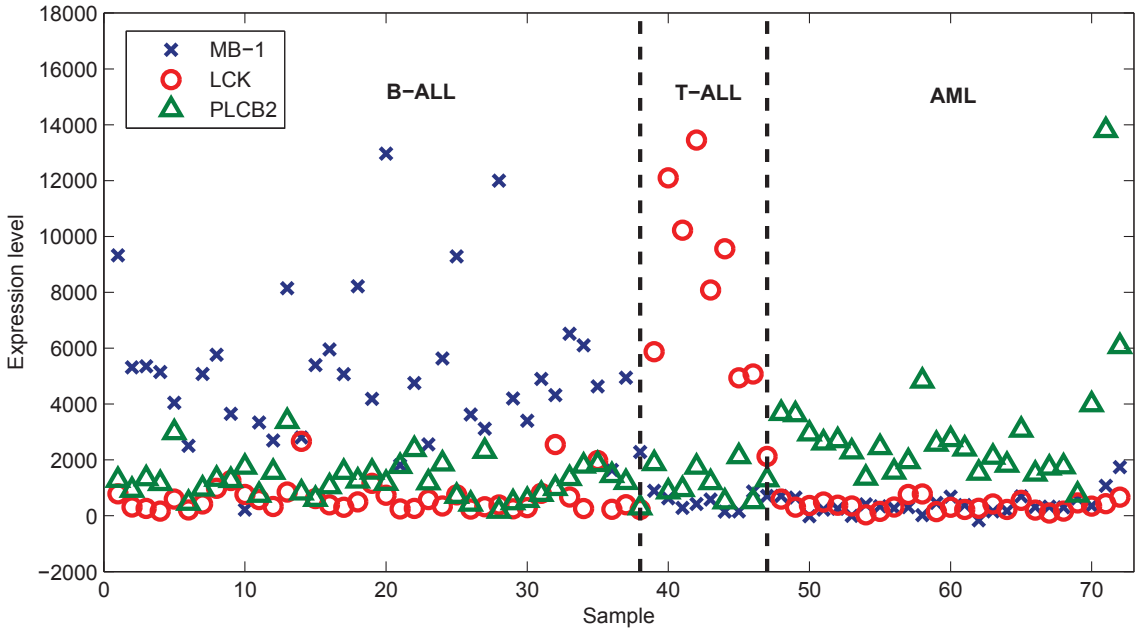

Figure 2.2: Gene expression pattern for a top scoring set.

ALL), with 25, 38 and 9 samples respectively. Figure 2.2 depicts a top scoring set that has been found consisting of gene *PLCB2*, *MB-1* and *LCK*. Class prediction for a particular sample is determined by the gene whose expression level in this set is the highest. As shown later, TSS yields 95.83% prediction accuracy on this data set using leave-one-out cross-validation.

In general, TSS searches for gene sets exhibiting a particular pattern that may be suitable for classification. There are, of course, many other patterns that one might consider with the potential for effective classification. Still, it is important to be mindful that increasing the size of the pattern search space would result in increases in already substantial computational costs, and is more likely to produce over-fitting.

Gene set score

To illustrate the scoring method, the example in Figure 2.2 is considered, and the goal is to score the gene set consisting of *PLCB2*, *MB-1* and *LCK*. Denote three genes as i_1,

i_2 and i_3 respectively, and the observed class conditional frequencies of their expression comparisons can be calculated in Table 2.1.

Table 2.1: Observed frequencies of expression comparison for a three-gene set.

	Leukemia		
	AML	B-ALL	T-ALL
$X_{i_1} > \max(X_{i_2}, X_{i_3})$	1	0	0
$X_{i_2} > \max(X_{i_1}, X_{i_3})$	0	0.9737	0
$X_{i_3} > \max(X_{i_1}, X_{i_2})$	0	0.0263	1

Interestingly, for AML, gene i_1 has the highest expression level among three genes in 100% of the samples, as indicated by the first column of Table 2.1. Similarly, for B-ALL, gene i_2 has the highest expression level in 97.4% of the samples, and for T-ALL gene i_3 has the highest expression level for 100% of the samples. Therefore, based on the information provided by these three genes, a natural way to classify a sample with expression levels x, would be to predict AML, B-ALL, and T-ALL respectively by determining which of the three expression levels x_{i_1}, x_{i_2}, and x_{i_3} is highest. Accordingly, we define a score for $\{i_1, i_2, i_3\}$ based on Table 2.1 as the sum of the row maxima, i.e.

$$\max\{1, 0, 0\} + \max\{0, 0.9737, 0\} + \max\{0, 0.0263, 1\} = 2.9737.$$

If the score is 3, clearly, the rule described above obtains a zero apparent error rate on the data set. Furthermore, if the underlying probability distributions of gene expression comparisons are well reflected by the observed frequencies, a higher score indicates that the rule is more likely to be effective for new samples. Therefore, the ultimate goal is to search for three-gene sets with the highest possible score.

In general, for an m-class problem, each m-gene set $\mathcal{S} = \{i_1, i_2, ..., i_m\}$ leads to a similar table as follows

Table 2.2: Observed frequencies of expression comparison associated with $\mathcal{S}$.

	Class			
	$y = 1$	$y = 2$	...	$y = m$
$X_{i_1} > \max\{X_r, r \in \mathcal{S} \setminus i_1\}$	$\hat{p}_{11}$	$\hat{p}_{12}$	...	$\hat{p}_{1m}$
$X_{i_2} > \max\{X_r, r \in \mathcal{S} \setminus i_2\}$	$\hat{p}_{21}$	$\hat{p}_{22}$	...	$\hat{p}_{2m}$
.....		.....		
$X_{i_m} > \max\{X_r, r \in \mathcal{S} \setminus i_m\}$	$\hat{p}_{m1}$	$\hat{p}_{m2}$	...	$\hat{p}_{mm}$

where $\hat{p}_{rj}$ is the frequency that given class $y = j$, gene i_r has the highest expression level in $\mathcal{S}$. The score for $\mathcal{S}$ is then defined as

$$\sum_{r=1}^{m} \max_{j=1,2,..m} (\hat{p}_{rj}). \tag{2.2.3}$$

The formula (2.2.3) can have a Bayesian decision-theoretic interpretation, where a Bayes optimal rule is chosen among a set of possible decision rules by minimizing the Bayes risk. However, this Bayesian optimality only applies when the gene set used for classification has been determined. Otherwise, the Bayes classifier requires the knowledge of the joint probability distribution of genes and classes, which is a formidable task. Details of the interpretation and some possible extensions of (2.2.3) are discussed in Section 2.5.1.

In practice, for further breaking ties among gene sets with the highest score, a sec-

ondary score is also considered based on Table 2.2 as

$$\sum_{j=1}^{m}\sum_{r=1}^{m}(-\hat{p}_{rj})\ln\hat{p}_{rj}. \tag{2.2.4}$$

Here $\sum_{r=1}^{m}(-\hat{p}_{rj})\ln\hat{p}_{rj}$ is the entropy of the estimated (class conditional) distribution for the expression comparison on the j-th row of Table 2.2. The smaller the entropy is, the more "peaked" the estimated distribution is in a certain class, which can lead to a more accurate classification rule. As a result, the secondary score (2.2.4) is defined as the sum of entropies for all possible expression comparisons. Each top scoring set that is finally chosen is also required to minimize this secondary score.

Decision rule

For each top scoring set $\tilde{\mathcal{S}}$, the prediction for sample x is

$$\hat{y} = \underset{c=1,2,..m}{\arg\max}\{x_{i_c}, i_c \in \tilde{\mathcal{S}}\}. \tag{2.2.5}$$

When $m = 2$, equation (2.2.5) turns out to be (2.2.1). Therefore, TSS is essentially a generalization of TSP in the multiclass case.

Although it rarely happens, due to expression level ties the decision rule for a single top scoring set can produce multiple classes associated with genes whose expression is the highest. In this situation, a randomized decision is introduced where each associated class is assigned a vote of $1/T$, where T is the number of genes producing the tie. For the final prediction, these votes are summed over the top scoring sets and the majority rule is applied.

Generally, finding the maximum element are one of the simplest types of analysis that can be drawn by comparing a set of elements. Because of this simplicity, the relation tends to be more robust against noise in the data, compared to any complex relations of

elements. For example, in the TST classifier [57], all possible permutations of a set of three genes are utilized to search for top scoring triplet(s) in a similar fashion of TSS. However, in principle, there are six possible permutations of their expression values, and the sample size needs to be large enough to estimate all class conditional probabilities associated with each permutation. The frequency estimates of a permutation can have a large variance if the sample size is small. Therefore, TST is less likely to be robust than TSS in the small-sample learning situation.

2.2.3 Greedy Search

In theory, TSS finds top scoring sets among all possible gene sets. For a data set with G genes and m classes, an exhaustive search has the complexity $O(G^m)$, which grows exponentially with respect to G. Hence, it is quite necessary to relax the global optimality requirement, which can be done in various ways. One idea would be to *a priori* reduce the search space. This can be typically done by pre-selecting a small number of genes based on a univariate multiclass criterion such as the one-way ANOVA F-test or the relative non-parametric Kruskal-Wallis test. Also, as presented in the next chapter, it is possible to use pathway information to restrict the search within naturally defined groups of genes. In this section, however, a different idea is proposed that adopts a greedy search algorithm to select gene sets that are top scoring in each of several stages, leading to what could be called *locally optimal* scoring sets.

For an m-class problem, the greedy search algorithm takes $m-1$ steps to form m-gene sets that are used in the final decision rule. It is initialized by finding the collection of gene pairs with the highest score for each of possible $\binom{m}{2}$ two-class (1-vs-1) sub-problems. Next, each possible two-class sub-problem is augmented by a single class, and for every

such augmentation, a collection of three-gene sets with the highest score is found based on top scoring gene pairs obtained in that two-class sub-problem. In particular, a distinct gene is added to each of such gene pairs to yield a group of three-gene sets, among which the ones with the highest score are sought. Then, the algorithm is performed iteratively until the size of sub-problems reaches m, and a collection of m-gene sets with the highest score is obtained from each sub-problem of size m. Finally, the (locally optimal) top scoring sets are found among all such collections for building a TSS classifier.

The greedy search process is illustrated in Figure 2.3. Since the first step involves two classes and each subsequent step deals with one more class, the formation of an m-gene set requires $m - 1$ steps. Importantly, all possible sequences of classification problems that start with a two-class problem and augmenting by one class at a time until reaching m classes are considered, so that ultimately, there are $m!/2$ collections of m-gene sets to be compared. The complexity of the first step is $O(G^2)$, and each following step only requires $O(G \cdot l)$ additional computations where l is the maximum size of collections of gene sets with highest scores generated in the previous step. Because the number such sets for a certain sub-problem is expected to be small, l is typically small so that the fully implemented algorithm has $O(G^2)$ complexity, which is significantly lower than the $O(G^m)$ complexity of an exhaustive search for $m > 2$.

The TSS classifier built by the greedy algorithm is typically validated by cross-validation. Normally, the greedy algorithm is assumed to be performed in each iteration of the cross-validation loop, which can lead to relatively extensive computation. To address this difficulty, an acceleration algorithm has also been introduced (see Section 2.4.2) that extends the pruning algorithm introduced by [83] for the TSP classifier to the multiclass case. In principle, the acceleration algorithm applies the greedy search method

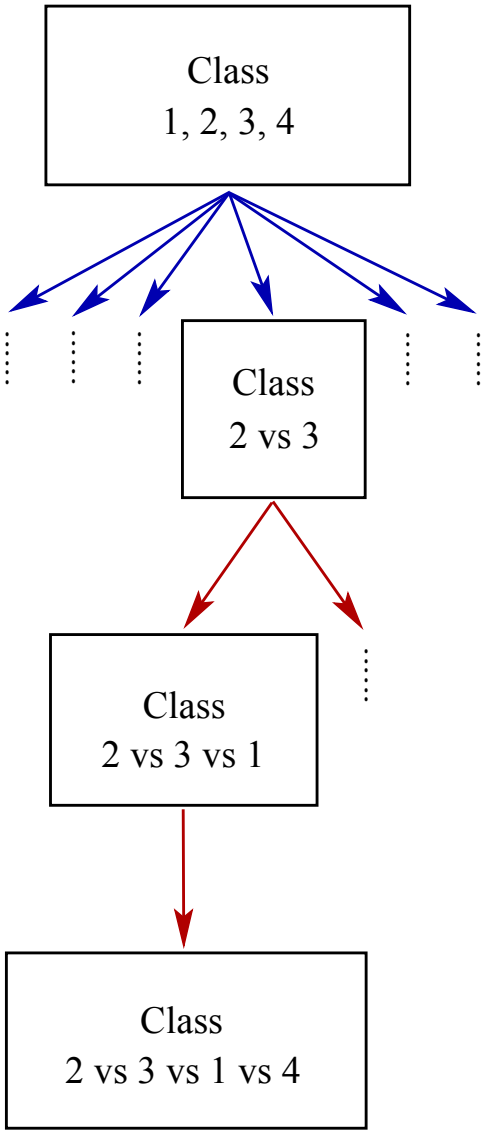

Figure 2.3: Schematic diagram of the greedy search algorithm. The workflow of the algorithm is illustrated for a four-class problem. Blue arrows represent the initialization step where each possible two-class sub-problems are considered. Each red arrow denotes an augmentation of the current problem by a single class. The graph shows one possible sequence of augmentations.

only one time on the entire data set and generates a small list of gene sets. Then, the top scoring sets identified from the list in each iteration of the cross-validation are guaranteed to be the same as those obtained by applying greedy search on the reduced training set.

The greedy algorithm is not likely to find gene sets having the globally maximum score. However since it is based on iterating through gene sets with highest scores in each possible sub-classification problem, the resulting efficiency gain appears substantial

enough to compensate for the fact that the space in which the search is carried out is limited, and produces high-scoring gene sets for the original problem. Also, since all possible sizes of sub-problems are investigated sequentially and only gene sets with the highest score are kept in each iteration, the final top scoring sets can be globally optimal if the global solution is also optimal in each of these sub-problems. In particular, this happens when the top scoring sets obtained by the algorithm have a perfect score.

2.2.4 Error Estimation

Error estimation always plays an crucial role in evaluation of classification methods, and it is frequently used as the guidance for selecting the best approach. The simplest metric often considered is misclassification rate (or equivalently classification accuracy). Its calculation is very efficient and immediately applies to any number of classes. One of the main drawbacks of misclassification rate is that it only indicates the performance of a trained classifier at a single operating point, which is controlled by the classification threshold used. In view of this limitation, a popular criterion preferred is the receiver operator characteristic (ROC) analysis. This analysis constructs an ROC curve by plotting sensitivity (true positive rate) versus one minus specificity (false positive rate), varying the decision threshold over its whole range. ROC analysis typically allows a classifier to be inspected over a range of possible conditions, and produces a scalar as the performance measure known as "the area under the ROC curve" (AUC). The AUC summarizes model performance over all possible thresholds, and the evaluation of classifiers based on AUC is typically independent of class priors, misclassification costs, and operating points. Therefore, AUC-based metrics have been extensively used in many areas. The original AUC measure is however only applicable to the two-class case. Hence, there are many attempts

for extending this popular metric in the multiclass setting. The approach discussed here is the one proposed by Hand and Till [40]. They suggested a simple extension of using binary AUC values to calculate the AUC for a K-class problem as

$$M = \frac{1}{K(K-1)} \sum_{i \neq j} \hat{A}(i|j).$$

M denotes the multiclass AUC, and $\hat{A}(i|j)$ is the estimated AUC of class i and j, or more formally, the probability that a randomly drawn member of class j will have a lower estimated probability of belonging to class i than a randomly drawn member of class i. This approach simply averages the AUCs resulting from $K(K-1)$ pairwise class comparisons (note that $\hat{A}(i|j) \neq \hat{A}(j|i)$).

The misclassification rate and the multiclass AUC are both considered for model evaluation in this thesis. In order to apply them to real data, some sampling strategies are often needed. There are two commonly used techniques. The first one is the "train-test split" scheme. The data are partitioned (either naturally or randomly) into a training and a test set. The training set is used for model optimization and fitting. The test set is for assessing the performance of the final chosen model. The second method is the "cross-validation" scheme. The data are partitioned into several roughly equal-sized folds. These folds are chosen as the test set one at a time, and each time the training set is selected as the data excluded the test set. A particular form of cross-validation employed throughout this thesis is leave-one-out cross-validation (LOOCV). If the data has N samples, LOOCV is a N-step loop. At each step, one distinct sample is left out while a classifier is trained on the remaining data and is tested the excluded sample. The final evaluation is based on N prediction results.

2.2.5 Ensemble Classification

Ensemble classification refers to the strategy of combining decisions of multiple classification models to reach a final prediction, and is often used for achieving a better classification performance. Ensemble methods are quite popular in general machine learning or pattern recognition problems, and they have also been applied to the gene expression domain [17, 83]. In principle, TSS is analogous to a decision stump (one-level decision tree), a simple decision rule that is flexible to be combined with any ensemble strategy. In this section, three main approaches are introduced to ensemble TSS for classification.

The most straightforward approach to ensemble classification when there are several classifiers is the majority voting. To implement this in the TSS method, instead of searching for the gene set with the highest score, those corresponding to the top k scores can be collected for majority voting. The final decision is the class that receives most votes, and ties, if any, are broken at random. A similar version of this ensemble approach has been adopted for TSP in the k-TSP method [83]. The only problem remaining here is the method for choosing k, which can be subjective. Re-sampling procedures for determing k such as cross-validation or Bootstrapping are often considered.

Majority voting assigns equal weights to all decision rules in the ensemble. An improvement might be achieved if more appropriate weights can be computed. In this instance, boosting is a classic example. It was first introduced by Freund and Schapire [29], with their AdaBoost algorithm for the two-class classification problem, and has been regarded as one of the most powerful techniques in supervised learning. An extension of AdaBoost, known as "SAMME", to the multiclass case was introduced by Zhu et al. [95]. Table 2.3 provides a brief description of SAMME when combined with TSS. Like the majority voting scheme, the number of boosting steps M here has to be determined.

Table 2.3: Description of the "SAMME-TSS" algorithm.

SAMME-TSS

Input: Training set $L = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ with m classes.

Output: Decision rule $C(x)$ for a sample x to be classified.

Algorithm:

1. Initialize the weights for samples $\omega_i = 1/N, i = 1, 2, ..., N$.

2. For $j = 1$ to M:

(a) Fit a TSS classifier $T_L^{(j)}$ to the training data using weights ω_i.

(b) Compute $err^{(j)} = \sum_{i=1}^{N} \omega_i \cdot \mathbb{I}_{\{y_i \neq T_L^{(j)}(x_i)\}}$.

(c) Compute $\alpha^{(j)} = \log \dfrac{1 - err^{(j)}}{err^{(j)}} + \log(m-1)$.

(d) Update N weights as $\omega_i \cdot \exp(\alpha^{(j)} \cdot \mathbb{I}_{\{y_i \neq T_L^{(j)}(x_i)\}})$.

(e) Re-normalize ω_i.

3. Output $C(x) = \arg\max_k \sum_{j=1}^{M} \alpha^{(j)} \cdot \mathbb{I}_{\{T_L^{(j)}(x_i)=k\}}$

The last ensemble approach discussed here is random forests. Random forests construct a number of decision trees and produces the final prediction through a majority vote by individual trees. In the well-known algorithm of random forests developed by Breiman [7], each of individual trees is fully grown but each node split in the tree is based on a small and random collection of features. If TSS is viewed as one level decision tree, the same idea can be used to ensemble TSS. To be specific, a multitude of TSS classifiers are constructed using the training data. Each classifier is built through an exhaustive search (for the top scoring set) in a few randomly chosen features. The final prediction is the mode of the classes selected by individual classifiers. The number of randomly chosen features is often pre-specified and is much less than the actual number of features. Therefore, this scheme can be very efficient.

Ensemble methods usually are able to achieve improved performance, but they tend to decrease the transparency and interpretability of the final decision rule. However, the failure to provide an interpretation does not necessarily diminish the potential clinical usefulness of biomarkers with unknown biological functions. Therefore, ensemble classification is tested in this chapter for determining potential improvements in the predictive performance of TSS on real microarray data.

2.3 Code Implementation

The TSS approach has been implemented using R 3.0.0 (http://www.r-project.org/) and C++. The interface between R and C++ is provided by the R package Rcpp (http://cran.r-project.org/web/packages/Rcpp/index.html). Besides the classification approach, methods proposed in this thesis for classifier training such as the greedy search algorithm have also been implemented in R. In this section, detailed instructions for using the corresponding R codes (http://jshare.johnshopkins.edu/dnaiman1/public_html/tss) are provided, including the configuration for the R environment, the data format and parameters for major functions.

The first step is to install the R software and the package Rcpp. The exact instructions may vary on different operating systems. Here are the steps for installation on Windows 7 Pro x64:

1. Download R (http://cran.r-project.org/bin/windows/base/) and run the installation.

2. Install Rtools (http://cran.r-project.org/bin/windows/Rtools/) that has the tool chain required for C++ code compilation.

3. Download batchfiles (http://cran.r-project.org/contrib/extra/batchfiles/) to always point to the latest version of R on the system when running R from the command line.

4. Download the Redmond Path Utility (http://download.cnet.com/Redmond-Path/3000-2094_4-10811594.html) to alter PATH variables of the system.

5. Edit the PATH variable to allow system wide access to the current version of R on the computer and components of Rtools.

6. Restart the system. Open R and install Rcpp by running the command: `install.packages(''Rcpp'')`.

More details can be found on the website: http://www.r-bloggers.com/installing-rcpp-on-windows-7-for-r-and-c-integration/. After successfully installing Rcpp, the next step is to load Rcpp using the `library` (or `require`) command in R:

```
> library(''Rcpp'')
```

Once Rcpp is loaded, the primary function for the TSS approach can be acquired by:

```
> sourceCpp(''TSS.cpp'')
```

Here assume all codes at http://jshare.johnshopkins.edu/dnaiman1/public_html/tss/R scripts/ are downloaded and placed in the same folder accessed by R.

Before introducing any specific function, users need to understand how to format the data. For use of the package, the data is assumed to take the form of a numerical matrix that contains expression levels from many microarray experiments where values in each row represent measurements for a particular gene across all patient samples and values in each column represent expression levels for a certain sample across all genes. This data format is very typical for functions handling gene expression data, and can

be created from scratch, loaded from external (e.g. Excel) files or extracted from widely used ExpressionSet objects in R. For example, to load the sample data set, the following R commands are used:

```
> trainData = read.csv('Leukemia1_train.csv',header=F,row.names=1)
> testData = read.csv('Leukemia1_test.csv',header=F, row.names=1)
> class = c(as.numeric(trainData[1,]),as.numeric(testData[1,]))
> trainData = as.matrix(trainData[-1,])
> testData = as.matrix(testData[-1,])
> dataM = cbind(trainData,testData)
```

In the codes above, two numerical matrices have been loaded as `trainData` and `testData`, and have been combined into a single matrix `dataM`. The class labels are stored in the first row of each matrix. Categorical classes need to be converted into numerical classes starting from 1. A summary of the data and class labels is obtained as:

```
> table(class)
class
1 2 3
38 9 25
> dim(dataM)
7129 72
```

The sample gene expression data contains microarray measurements for 7129 genes and 72 samples, which are grouped into 3 classes with 38, 9 and 25 samples respectively. Alternatively, the expression data can be obtained from ExpressionSet objects in R as follows:

```
> require(golubEsets)
```

```
> data(Golub_Train)
> data(Golub_Test)
> trainData = exprs(Golub_Train)
> testData = exprs(Golub_Test)
> dataM = cbind(trainData,testData)
```

After obtaining the data matrix and class labels, it is now possible to determine the TSS classifier. A simple example is as follows:

```
> geneIdx = c(1:10)
> geneSets = t(combn(geneIdx,3))
> tss = TopScoringSet(dataM,class,3,geneSets)
```

Here the sample data set contains 3 classes, so the TSS approach aims to find the top scoring 3-gene sets. `geneIdx` specifies the indices of genes of `dataM` used to form the collection of possible 3-gene sets. The `geneSets` above takes the form:

```
> head(geneSets)
1 2 3
1 2 4
1 2 5
1 2 6
1 2 7
1 2 8
```

In this example, the first 10 genes of `dataM` are used to form $\binom{10}{3}$ 3-gene sets, among which the `TopScoringSet` function returns the top scoring sets. In general, `TopScoringSet` takes a numerical matrix (expression data), a numerical vector (class labels) and an integer (number of classes) as arguments. Additionally, a numerical matrix

`geneSets` needs to be provided as the collection of gene sets considered where each row of the matrix corresponds to the indicies of a single gene set (as shown above).

The output of the `TopScoringSet` function is a list object that contains a number of sublists where each sublist includes all information for a single top scoring set. The output `tss` of the example above is:

```
> tss

$set1

$set1$gene

7 9 10


$set1$maxScore

1.312398


$set1$class

0 2 1


$set1$classScore

0.2368421 0.5200000 0.5555556
```

The result here indicates that there exists only one top scoring set consisting of the 7-th, 9-th and 10-th gene of `dataM`. The top score is 1.31 and the classes where each gene achieves the maximum expression level relative to the other two genes are contained in `class`. In addition, the frequency estimate of the "max" gene in each class is included in `classScore`.

The top scoring set in the example above has the score of 1.31, which is significantly

lower than the perfect score 3. This is mostly due to the restriction to use only the first 10 genes. In principle, `geneSets` has to include all possible 3-gene sets that can be formed from 7129 genes of `dataM`. However, as discussed in the "Greedy Search" section, this leads to an exhaustive search that has an undesirable computational complexity. Therefore, `geneSets` often needs to be pre-selected, which can be done in various ways. For example, the Kruskal-Wallis test can be used for this purpose:

```
> source('KWTest.r')
> geneIdx = KWTest(dataM,class,100)
> geneSets = t(combn(geneIdx,3))
> head(geneSets)
235 758 760
235 758 804
235 758 874
235 758 922
235 758 1078
235 758 1120
> tss = TopScoringSet(dataM,class,3,geneSets)
$set1
$set1$gene
758 2833 3433


$set1$maxScore
2.947368
```

```
$set1$class

0 1 2


$set1$classScore

0.9473684 1.0000000 1.0000000
```

In the codes above, the top 100 genes selected by the Kruskal-Wallis test are kept and all $\binom{100}{3}$ combinations form `geneSets`. The result then seems to be significantly improved by this pre-selection. The top scoring set is a single gene set consisting the 758-th, 2833-th and 3433-th gene of `dataM`, ane this set achieves a high score of 2.95. The later two genes have the maximum expression levels in 100% of samples in class 1 and 2 respectively.

The Kruskal-Wallis test works well as an efficient filtering method for the example here, but as discussed before, it may not perform well in general. One effective approach proposed in this thesis is the greedy search algorithm, which could address the limitations of common univariate tests as filtering methods for the TSS classifier. The codes for using this algorithm are given below:

```
> classM = t(combn(c(1:3),2))

> setList = GreedySearch_init(dataM,class,3,classM)

> classM

1 2

1 3

2 3

> setList = GreedySearch_augt(dataM,class,3,classM,setList)
```

The greedy search algorithm contains an initialization step and several augmentation

steps. The initialization step starts by working on each possible pair of classes, which are specified in `classM`. After the initialization step, each pair of classes is augmented by a new class and the process is repeated until no new class exists. The augmentation process needs only one function (`GreedySearch_augt`) call that loops over all augmentation steps. Since the sample data above has 3 classes, there is actually only one augmentation step carried out by the greedy search algorithm. The resulting `setList` is a list object consisting a number of sublists, each of which contains all candidate gene sets (for top scoring sets) for a particular augmentation path (i.e., sequence of augmentations, see Figure 2.3). For example, the `setList` object above has three sublists corresponding to three augmentation paths:

```
> dim(setList[[1]])
7 3
> dim(setList[[2]])
1 3
> dim(setList[[3]])
73 3
```

As shown above, these three sublists consist of 7, 1 and 73 3-gene sets respectively. In the final step of greedy search, gene sets obtained from all augmentation paths are collected as candidate gene sets. But duplicates need to be removed because no duplicates are allowed in the list of top scoring sets. The following codes are implemented for this purpose:

```
> geneSets = integer(0)
> for(i in 1:length(setList))
> geneSets = rbind(geneSets,setList[[i]])
```

```
> for(i in 1:dim(geneSets)[1])
> geneSets[i,] = sort(geneSets[i,])
> st = rep(1,dim(geneSets)[1])
> for(i in 2:dim(geneSets)[1])
> for(j in 1:(i-1))
> if(sum(geneSets[i,]==geneSets[j,])==3)
> st[i] = 0
> geneSets = geneSets[st==1,]
> dim(geneSets)
74 3
```

The ultimate collection contains 74 unique gene sets from greedy search, and a TSS classifier is built as follows:

```
> tss = TopScoringSet(dataM,class,3,geneSets)
> length(tss)
73
tss[[1]]
$gene
88 2642 4342


$maxScore
2.973684


$class
2 0 1
```

```
$classScore

1.0000000 0.9736842 1.0000000
```

The TSS classifier has 73 top scoring sets with the same top score 2.97. Compared to the result of using the Kruskal-Wallis test, the greedy search algorithm seems to work better by finding top scoring sets with higher score, which is also done by searching within fewer candidates. But it is important to be mindful that the greedy search algorithm is likely to need considerably more computational effort than common univariate tests. One such top scoring set consisting of the 88-th, 2642-th and 4342-th gene of `dataM` is illustrated above. Finally, the classifier `tss` can be used to predict the classes of unlabeled samples as follows:

```
> source('predict_TSS.r')

> predclass = predict_TSS(tss, testData)
```

The `predict_TSS` function takes the object returned by `TopScoringSet` and a numerical matrix (test set) as arguments. Note that `testData` here is assumed to have the same group of genes with the same indices as in the training set. As a result, `predclass` contains class predictions in the same format as for `class` used in the training process (sample IDs are shown):

```
> predclass

V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 3 1

V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35

2 1 1 3 2 3 2 1 3 3 1 2 1 1
```

2.4 Validation

2.4.1 Classification of Human Cancer Microarray Data

The TSS classifier was validated on seven gene expression microarray data sets retrieved from public databases or authors' websites (Table 2.4). They are related to human cancers including leukemia (Leukemia), mixed lineage leukemia (MLL), lung adenocarcinoma (Lung), small round blue cell tumors (SRBCT), bladder carcinoma (Bladder), childhood acute lymphoblastic leukemia (ChildALL) and non-small cell lung cancer (NSCLC). The types of problems range from classification of cancer subtypes (e.g., Leukemia, SRBCT and NSCLC), tumor stages (e.g., Lung and Bladder) and treatment responses (e.g., ChildALL). Data sets with GEO accession number are obtained from the Gene Expression Omnibus website (http://www.ncbi.nlm.nih.gov/geo/). Most data sets were produced using Affymetrix platforms with the exception that SRBCT was generated using a customized cDNA micoarray. Standard data preprocessing methods were applied. Some of these data sets have already been investigated previously [60, 83] for evaluating gene expression classifiers. Additional information can be obtained from the references included in Table 2.4.

The top scoring gene sets found on seven data sets are summarized in Table 2.5. For each data set, top scoring sets have been obtained by applying the greedy search algorithm on all samples. The corresponding top score and re-substitution error have also been calculated.

In summary, there are 73 top scoring sets for Leukemia, seven sets for MLL, three sets for Bladder, two sets for SRBCT and NSCLC, and one for Lung and ChildALL. Only a few genes are actually involved in these sets and most genes appear in multiple sets.

Table 2.4: Seven gene expression data sets used for evaluating classification performance.

Dataset	Platform	#Classes	#Genes	#Samples	Source	Reference
Leukemia	Affymetrix HuGeneFL	3	7129	72	Authors' website	[35]
MLL	Affymetrix HGU95	3	12582	72	Authors' website	[3]
Lung	Affymetrix HuGeneFL	3	7129	96	Authors' website	[5]
SRBCT	cDNA	4	2308	83	Authors' website	[85]
Bladder	Affymetrix HuGeneFL	3	7129	40	GEO: GSE89	[21]
ChildALL	Affymetrix HGU95	4	12625	60	GEO: GSE412	[9]
NSCLC	Affymetrix HGU95	3	12599	33	GEO: GSE1987	[13]

Table 2.5: Top scoring gene sets identified on seven gene expression data sets.

Dataset	No. of sets	No. of genes	Score	Error
Leukemia	73	36	2.97/3.00	1/72
MLL	7	12	2.96/3.00	1/72
Lung	1	3	2.72/3.00	16/96
SRBCT	2	5	3.93/4.00	2/83
Bladder	3	5	3.00/3.00	0/40
ChildALL	1	4	3.09/4.00	13/60
NSCLC	2	6	2.88/3.00	1/33

It is interesting to note that some of these genes would not be regarded as differentially expressed based on their individual expression values, but the relative comparison of expression levels in each top scoring set produces enhanced class separability. The observed frequencies of some top scoring sets are displayed in Table 2.6. For each set, the table gives the relative frequency at which the maximum expression value appears among genes in the set for every class. In each case, these relative frequencies provide good evidence

for discriminability of the set, which indicates the potential for class prediction.

Table 2.6: Observed frequencies of gene expression comparison in two top scoring sets from (a) NSCLC and (b) SRBCT respectively.

(a)

	Max gene		
Class	*KRT14*	*CNGB1*	*GDF10*
SCC	1	0	0
ADCA	0	1	0
N	0	0.12	0.88

(b)

	Max gene			
Class	*GYG2*	*EST*	*CDH2*	*HCLS1*
EWS	0.93	0	0.07	0
RMS	0	1	0	0
NB	0	0	1	0
BL	0	0	0	1

Next, to validate the greedy search-based TSS (G-TSS), its classification accuracy was assessed on seven microarray data sets. As a comparison to the greedy search algorithm, a common differential expression technique was considered based on the Mann-Whitney test and the "1-vs-all" strategy to select top n genes for separating each class from the union of other classes. To save computation time, n was chosen to be 50 for three-class and 25 for four-class problems. The resulting TSS classifier (denoted as "MW-TSS") was compared to G-TSS in terms of classification accuracy. Furthermore, five popular machine learning techniques were also considered as benchmarks to the TSS approach: k-nearest

neighbors (kNN), naive Bayes (NB), random forests (RF), support vector machines with a linear kernel (l-SVM) and PAM. No feature selection algorithm is applied to benchmark methods. All analyses were performed using packages in R 3.0.0. LIBSVM [10] was used as the implementation for SVMs. There are a variety of model choices provided by LIBSVM and the linear kernel SVM is suggested for microarray data. In particular, multiclass problems are handled by LIBSVM using the "1-vs-1" approach.

The performance of a classifier was measured by classification accuracy through leave-one-out cross-validation (LOOCV). For classifiers with parameters (e.g., the number of nearest neighbors k in kNN and the cost factor C in l-SVM), the performance evaluation was realized by a double LOOCV loop. To be precise, a double LOOCV loop consists of an inner loop and an outer loop, both of which use the leave-one-out partition scheme. For each classifier model, the inner loop is responsible for model optimization that usually involves parameter tuning, and the outer loop is used for calculating accuracy by averaging classification results. To avoid any optimistic evaluation result, each step of the outer loop is accomplished so that the training data on which the model optimization is performed is totally independent of the left out testing sample.

Table 2.7 provides a comparison of classification accuracies for different methods on seven data sets. In general, G-TSS has achieved comparable or better performance on most data sets. For Leukemia and MLL, it competes with the highest accuracies obtained by PAM and NB respectively. For Bladder, ChildALL and NSCLC, it turns out to be the most accurate classifier. In contrast, MW-TSS only yields comparable results on Leukemia and NSCLC and has the lowest accuracies on four out of seven data sets, which seems to indicate the inappropriateness of using traditional differential expression methods that focus on individual expression values to search for relative expression patterns.

Table 2.7: Comparison of classification accuracies estimated using LOOCV. The highest accuracy for each data set is highlighted in boldface.

Method	Leukemia	MLL	Lung	SRBCT	Bladder	ChildALL	NSCLC
G-TSS	95.83	94.44	70.83	90.36	**100.00**	**48.33**	**90.91**
MW-TSS	95.83	76.39	62.50	89.15	57.50	36.67	87.87
kNN	81.94	91.67	75.00	95.18	77.50	45.00	72.72
NB	94.44	**95.83**	75.00	98.80	82.50	46.67	66.67
RF	93.06	94.44	78.13	**100.00**	90.00	**48.33**	72.72
PAM	**97.22**	93.06	70.83	**100.00**	85.00	31.67	69.70
l-SVM	93.06	94.44	**83.33**	**100.00**	92.50	41.67	81.82

These results demonstrate the superiority of the greedy search algorithm for building the TSS classifier.

In this study, kNN, NB, RF and l-SVM make use of all available genes for classification. Although their performances could often be improved by feature selection or ensemble approaches, investigation of such improvements is beyond the scope of this thesis as the goal is not to merely develop a more accurate classifier. Instead, the competitive performance of TSS across all data sets demonstrates its stability. More importantly, the TSS approach is able to discover small informative subsets of genes, and its decision rule, compared with those of benchmark classifiers, proves to be much simpler, hence is more likely to provide for improved biological interpretability without a concomitant sacrifice in performance.

A Large Sample Case

While the predictive ability of the TSS classifier has been demonstrated across seven gene expression data sets, these data sets generally have a very limited sample size and

a small number of classes. To address these limitations, an attempt was made to apply the TSS approach to one extreme case: the Microarray Innovations in Leukemia (MILE) study program [37]. MILE is claimed as one of the largest gene expression microarray profiling studies in hematology and oncology. The expression profiles are collected from 11 laboratories in seven countries across three continents and consist of leukemia subtypes of myeloid and lymphoid malignancies. MILE is a two-stage study where a retrospective stage I generated expression profiles for 2,143 patients and was designed for biomarker discovery. A prospective stage II produced an independent cohort of 1,152 patients and was used for validation. Stage I used commercially available whole-genome microarrays (Affymetrix HG-U133 Plus 2.0) and stage II was performed using a newly designed custom chip (Roche AmpliChip). The microarray data have been deposited in Gene Expression Omnibus database under series accession number GSE13204.

MILE provides an unique opportunity for validating microarray-based classification models, especially for multiclass approaches. Each of 2,143 samples in stage I contains 54,675 gene expression measurements (45 missing values). Samples are classified into 18 diagnostic *gold standard* categories including eight ALL subtypes, six AML subtypes, two chronic leukemia subtypes, myelodysplastic syndromes and normal bone marrow. Stage II contains only 1,480 (1,457 disease-related and 23 housekeeping) genes and samples are also classified into 18 classes as defined in stage I. In the initial MILE study, a classification model was trained and tested for distinguishing all 18 classes. The multiclass model consists of binary classifiers formed by support vector machines with a linear kernel (l-SVM), each of which separates a pair of classes. In the results, high accuracies were observed for most classes, indicating the robustness of microarray-based classification. To compare the predictive performance, a classification model was trained using the TSS

approach on stage I samples and tested independently on samples from stage II. Since in the original MILE paper, the independent validation results were only shown for an acute leukemia diagnostic classifier, all 14 acute leukemia subtypes (Table 2.8) are considered here. Also, both training and test set contain only 1,457 genes that are in common for microarray data sets from two stages.

Although in thoery the TSS approach can be applied to any number of classes, the predictive power of top scoring sets through relative comparison is expected to decrease as the number of classes under study increases. Also, since the class hierarchy of acute leukemia subtypes is known *a priori*, it is better to incorporate such biological information to guide the decision process. Therefore, for this large multiclass problem, a two-step decision tree as shown in Figure 2.4 was introduced based on three TSS classifiers. The hierarchy of the tree is generated from the structure of the data where 14 acute leukemia subtypes can be grouped into three major lineage leukemias (B-ALL, T-ALL and AML). The B-ALL class is further divided into seven subtypes while the AML class contains six subtypes. As a result, three TSS classifiers were built for a three-class, six-class and seven-class problem respectively. The final prediction for a sample would follow the decision tree. In addition, to further improve the predictive performance, a similar procedure as introduced by [83] was considered to construct an ensemble of TSS classifiers for each of three multiclass problems. Specifically, the top k scoring gene sets were selected at each step of the greedy search process and the final prediction was the class receiving the majority of votes from k chosen gene sets. k was considered as the model parameter and the best $k \in \{1, 2, ..., 50\}$ was determined by LOOCV on the training set. Also, the acceleration algorithm was used to expedite the cross-validation process (see Section 2.4.2).

Table 2.8: Samples of acute leukemia subtypes used for classification. Three major leukemia classes consist of 14 subtypes. The class labels (C1 to C14) are the same as defined in the MILE study.

Class	Diagnosis	No. of samples	
		Training	Test
-	B-ALL	576	357
C1	Mature B-ALL with t(8;14)	13	5
C2	Pro-B-ALL with t(11q23)/*MLL*	70	23
C3	c-ALL/Pre-B-ALL with t(9;22)	122	62
C5	ALL with t(12;21)	58	64
C6	ALL with t(1;19)	36	10
C7	ALL with hyperdiploid karyotype	40	35
C8	c-ALL/Pre-B-ALL without t(9;22)	237	158
C4	T-ALL	174	79
-	AML	542	257
C9	AML with t(8;21)	40	16
C10	AML with t(15;17)	37	20
C11	AML with inv(16)/t(16;16)	28	20
C12	AML with t(11q23)/*MLL*	38	17
C13	AML with normal kt./other abn.	351	160
C14	AML complex aberrant karyotype	48	24

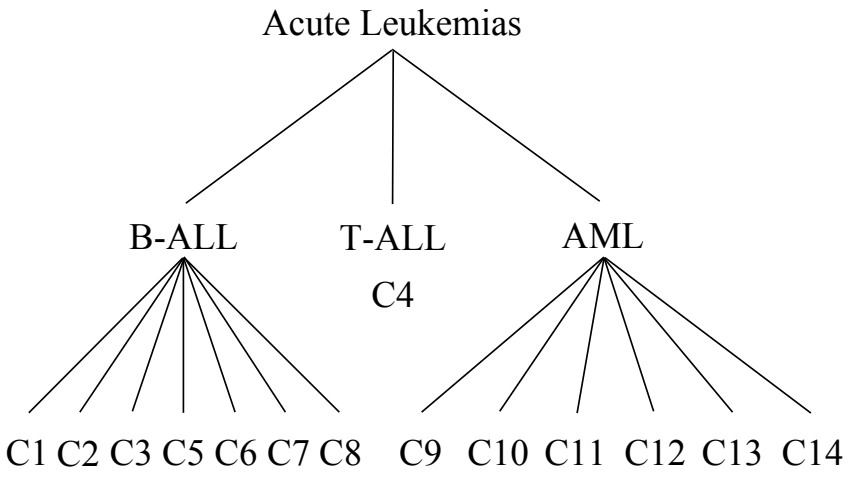

Figure 2.4: Two-step decision tree for classification of acute leukemia samples.

Prediction accuracies of G-TSS are shown in Table 2.9 and are compared to those achieved by l-SVM as presented in [37]. The optimal k for the ensemble of TSS classifiers in the three-, six- and seven-class problem (Figure 2.4) is 12, 7 and 42 respectively. G-TSS has achieved 100% correct predictions for two classes (C1 and C6), and $> 90\%$ accuracies for three classes (C4, C5 and C10). It outperforms l-SVM in three classes, and yields equal results in four classes. In general, there are at least 10 out of 14 classes in which comparable accuracies have been obeserved for both methods. For G-TSS, low accuracies are observed for C8, C12 and C13. For C12, its intrinsically heterogeneous nature has been discussed in [37]. For C8 and C13, this could be due to the imbalanced training sample sizes that may violate the equal prior probabilities for classes assumed by TSS. Nonetheless, G-TSS uses only three TSS classifiers as compared to the multiclass model that contains $\binom{14}{2} = 91$ SVM classifiers in the MILE study. Many fewer genes are hence involved in making predictions through G-TSS.

For a more comprehensive comparison, the two-step decision tree was also considered to combine with l-SVM (denote this approach as Hierarchical SVM or HC-SVM) where three l-SVM classifiers were trained for three multiclass problems. In each problem,

HC-SVM still used the one-vs-one scheme to handle multiple classes and the parameter tuning was realized using 10-fold cross-validation on the training set. Table 2.9 shows that HC-SVM has achieved similar accuracies to G-TSS and l-SVM in most classes. For C3, C5 and C14, the accuracy of HC-SVM is significantly lower than that of G-TSS and l-SVM, but good results have been observed in C8 and C13. Overall, the comparison demonstrates the effectiveness of G-TSS on a large data set. Its classification performance is as powerful as that of advanced machine learning classifiers while its decision boundaries remain transparent and potentially interpretable.

2.4.2 Cross-Study Comparison of Bladder Cancer

Bladder cancer is a common malignant disease that causes 145,000 deaths worldwide annually [22]. According to American Cancer Society (www.cancer.org/), the stage of disease at diagnosis is one of the most important factors in choosing treatment options and predicting a person's prognosis. A staging system is a standard way to describe the extent of cancer spread. For bladder cancer, a common staging system used is the "T" system where the letter T is followed by numbers and/or letters to describe how far the primary tumor has grown through the bladder wall and whether it has grown into nearby tissues. Higher T numbers mean more extensive growth. Bladder cancer typically progresses from stage Ta (non-invasive papillary carcinoma) or Tis (non-invasive flat carcinoma) to T1 (grown into the connective tissue below), T2 (grown into the muscle layer), T3 (grown into the fatty tissue layer) and T4 (spread into nearby organs or structures). Non-muscle-invasive tumors (Ta and T1) and muscle-invasive tumors (T2-T4) differ significantly in clinical treatment. Hence, there are many attempts for identifying prognostic variables of bladder cancer stages, and gene expression-based approaches have

Table 2.9: Comparison of acute leukemia classification methods. The number of correct classifications is followed by the corresponding accuracy (in percentage) for each class.

Class	G-TSS	l-SVM	HC-SVM
C1	5 (100.0)	4 (80.0)	4 (80.0)
C2	20 (87.0)	23 (100.0)	21 (91.3)
C3	51 (82.3)	53 (85.5)	34 (54.8)
C4	75 (94.9)	75 (94.9)	74 (93.7)
C5	62 (96.9)	59 (92.2)	13 (20.3)
C6	10 (100.0)	10 (100.0)	10 (100.0)
C7	30 (85.7)	22 (62.9)	32 (91.4)
C8	76 (48.1)	141 (89.2)	138 (87.3)
C9	14 (87.5)	16 (100.0)	16 (100.0)
C10	19 (95.0)	19 (95.0)	18 (90.0)
C11	17 (85.0)	20 (100.0)	16 (80.0)
C12	11 (64.7)	15 (88.2)	15 (88.2)
C13	127 (79.4)	148 (92.5)	156 (97.5)
C14	17 (70.8)	17 (70.8)	9 (37.5)

also been investigated. Dyrskjot et al. [22] conducted one of the largest cohort studies of bladder cancer. Microarray profiles of 404 patient samples diagnosed with bladder cancer were collected in hospitals in Denmark, Sweden, France, England, and Spain. A variety of classification tasks were performed using gene expression with respect to stage, recurrence, carcinoma *in situ* and progression to validate the prognostic value of molec-

ular classifiers. In this section, the classification of cancer stages is focused using the same microarray data, and the TSS classifier is considered to combine several ensemble methods.

Leave-Study-Out Cross-Validation

The bladder cancer data in [22] consists of patient samples collected independently from five locations. Since the batch/study effects of microarray data are widely observed [54], microarray classifiers are expected to generalize well on an independent test set. For this purpose, data sets from different platforms or locations are often integrated to demonstrate the robustness of a certain classifier. Specifically, in this section, classifiers are validated using a procedure called leave-study-out cross-validation. Each iteration of this procedure reserves samples from one location for testing and uses all other ones for training. The results from all iterations are then aggregated to estimate classification performance.

Table 2.10 summarizes the information of bladder cancer patients. The microarray data is publicly available using GEO (Gene Expression Omnibus) accession number GSE5479. Gene expression profiling was assayed on custom cDNA microarray. Preprocessing methods applied include the Lowess normalization [71] and a log 2 ratio transformation. The final data contains measurements of 1381 genes. Patient samples are from stage Ta, Tis, T1, T2, T3 and T4, but 10 samples from stage Tis, T3 and T4 are excluded in this study because of the small sample size. Therefore, the stage classification considered here is a three-class problem.

Classification performance was evaluated using the multiclass ROC analysis introduced in the "Error Estimation" section. The AUC of a classifier was obtained using

Table 2.10: Clinical information of bladder cancer patients.

	Spain	France	Denmark	Sweden	England	Total
Stage						
Ta	34	25	86	26	17	188
T1	7	19	63	75	8	162
T2	14	17	4	9	–	44

the train-test split strategy where leave-study-out cross-validation was used to partition the data. Patient samples from England were only considered for training because they have no samples for stage T2. The G-TSS (greedy search-based TSS) approach was applied and its AUC was compared with that of four benchmark classifiers including naive Bayes (NB), random forest (RF), PAM and support vector machine with a linear kernel (l-SVM). Parameters were tuned using 10-fold cross-validation on the training set.

Table 2.11: Comparison of AUCs across five locations.

	Spain	France	Denmark	Sweden	Average
NB	64.35	82.93	81.79	68.55	74.41
RF	71.35	80.98	86.40	68.58	76.83
PAM	71.22	76.50	86.96	74.31	77.25
l-SVM	72.95	78.58	83.54	62.54	74.40
G-TSS	70.27	74.4	70.41	58.23	68.33

Results in Table 2.11 demonstrate that AUCs of G-TSS across five locations are consistently lower than those of any benchmark method, although it is not the worst classifier in terms of misclassification rates (not shown). This is partially due to the fact that the calculation of AUC is based on estimates of posterior probabilities used by

decision rules. The rules of benchmark classifiers typically assign non-zero probabilities to all classes with the highest probability to the most desirable class, which can benefit from wrong decisions since there is still a non-zero probability of choosing the "true" class. In contrast, the rule of TSS is often based on a single top scoring set, and the possible decision for a class then can only be "yes" or "no". At a particular decision threshold (e.g., misclassification rate), TSS seems to achieve better rates, but it can be less robust in a broad sense.

TSS Ensembles

Motivated by the limitation discussed above, an attempt was made to combine TSS with ensemble methods described in the "Ensemble Classification" section. In particular, top scoring gene sets were utilized as "weak" classifiers (classifiers that might not perform much better than random guessing) and were combined by some voting scheme into a "powerful" classifier. Specifically, three ensemble schemes were considered and compared: "k-TSS" collects top k scoring sets obtained by the greedy search algorithm and allows them to make a majority vote for the final decision; "RF-TSS" imitates the process of random forest to build a number of TSS classifiers, each of which is based on a small group of randomly selected genes. Again the majority rule is used to reach the final prediction; "SAMME-TSS" adopts the "SAMME" algorithm, a multiclass ensemble strategy [95] that extends the well-known Adaboost method in the multiclass case. Basically, SAMME-TSS combines gene sets with top scores for prediction with appropriately chosen weights, which are obtained sequentially by minimizing some pre-defined cost function. Three ensemble methods were used to improve the classification results.

AUC values at various ensemble sizes for k-TSS are shown in Table 2.12. AUCs were

calculated for four locations (Spain, France, Denmark and Sweden) whose samples were tested. These results show that significant improvements in terms of AUC have been obtained by k-TSS.

Table 2.12: AUCs of k-TSS at various ensemble sizes across different locations.

Ensemble Size	Spain	France	Denmark	England	Average
No ensemble	70.27	74.4	70.41	58.23	70.27
10	72.62	80.33	80.62	67.88	75.36
50	73.47	80.59	83.08	68.45	76.40
100	73.42	84.41	83.98	69.34	77.79

Also, the AUC values of RF-TSS were achieved in Table 2.13. To be precise, each of many TSS classifiers in RF-TSS was built through an exhaustive search among a number of randomly chosen genes. Similar to the random forest setting, the number was selected as $\sqrt{1381} \approx 37$ (square root of the number of genes), and various ensemble sizes were used for training.

Table 2.13: AUCs of RF-TSS for different ensemble sizes.

Ensemble Size	Spain	France	Denmark	England	Average
No ensemble	70.27	74.4	70.41	58.23	70.27
10	63.86	67.84	68.12	46.26	61.52
50	65.55	76.63	73.97	56.63	68.20
100	63.45	81.74	75.28	57.35	69.46

AUCs in Table 2.13 show little improvement after using RF-TSS. Some were even worse for small ensemble sizes. Different random selections were tried and the results seemed to be similar. The performance of RF-TSS is likely to be degraded by "noisy"

TSS classifiers in the ensemble. In contrast, k-TSS obtains better results with many fewer gene sets.

The last method investigated is SAMME-TSS. The algorithm is unambiguously detailed in Table 2.3 except for step 2(a) in which a TSS classifier is built on the weighted training data. Typically this step is realized by searching for a weak classifier that minimizes the weighted error rate. In SAMME-TSS, the set of all weak classifiers was the collection of top k scoring sets generated by the greedy search algorithm and the minimization was acutally carried out over k elements. The value of k was chosen to be large enough ($k = 10,000$) so that there were enough weak classifiers to choose from.

Table 2.14: AUCs of SAMME-TSS for different ensemble sizes.

Ensemble Size	Spain	France	Denmark	England	Average
No ensemble	70.27	74.4	70.41	58.23	70.27
10	71.11	86.21	72.85	53.65	70.96
50	64.71	84.53	78.32	56.18	70.94
100	66.6	84.63	78.92	56.3	71.61

The AUCs of using SAMME-TSS were listed in Table 2.14. The results seem slightly better than RF-TSS, but are generally worse than k-TSS. The performance on some data (e.g. France) appears to be boosted, and the AUCs typically decrease when ensemble size increases. The boosting procedure could be limited by the set of pre-selected gene sets. Overall, k-TSS turns out to be the best ensemble approach for this bladder cancer cohort study.

2.5 Theoretical Results

2.5.1 Bayesian Decision-theoretic Interpretation

In section 2.2.2, the formula (2.2.3) was used to calculate the score of a gene set in the TSS classifier. This section gives a theoretical derivation of this formula and its form with a general loss function and class priors using the Bayesian decision theory.

For an m-class classification problem, each m-gene set $\mathcal{S} = \{i_1, i_2, ..., i_m\}$ selected by TSS is supposed to satisfy (2.2.2). Now assume the class conditional probability distributions associated with gene expression comparisons in $\mathcal{S}$ are given exactly in Table 2.15.

Table 2.15: Class conditional probabilities of expression comparisons associated with $\mathcal{S}$.

	Class			
	$y = 1$	$y = 2$	...	$y = m$
$X_{i_1} > \max\{X_r, r \in \mathcal{S} \setminus i_1\}$	p_{11}	p_{12}	...	p_{1m}
$X_{i_2} > \max\{X_r, r \in \mathcal{S} \setminus i_2\}$	p_{21}	p_{22}	...	p_{2m}
.....		.....		
$X_{i_m} > \max\{X_r, r \in \mathcal{S} \setminus i_m\}$	p_{m1}	p_{m2}	...	p_{mm}

Then a decision procedure δ for class prediction can be constructed where each comparison above is considered being indicative of a sample from a distinct class. Consequently, m comparisons lead to $m!$ possible δs for a given gene set, and one of them is illustrated below.

Next, a loss function can be introduced for δ by specifying the penalties for each type of misclassification as in Table 2.17 Based on the tables above, $R(i, \delta)$, the risk function

Table 2.16: Decision procedure δ based on m possible relations resulted from expression comparison of genes in $\mathcal{S}$.

X	$\delta(X)$
$X_{i_1} > \max\{X_r, r \in \mathcal{S} \setminus i_1\}$	3
$X_{i_2} > \max\{X_r, r \in \mathcal{S} \setminus i_2\}$	1
.....	.....
$X_{i_m} > \max\{X_r, r \in \mathcal{S} \setminus i_m\}$	2

Table 2.17: Loss function for decision procedure δ

	$\delta = 1$	$\delta = 2$	...	$\delta = m$
$y = 1$	l_{11}	l_{12}	...	l_{1m}
$y = 2$	l_{21}	l_{22}	...	l_{2m}
...	.....			
$y = m$	l_{m1}	l_{m2}	...	l_{mm}

of δ for class $y = i$ can be written as

$$R(i, \delta) = \sum_{j=1}^{m} l_{ij} \cdot p(\delta = j | y = i) \tag{2.5.1}$$

where $p(\delta = j|y = i)$ corresponds to one of probabilities in Table 2.15 and can be determined using Table 2.16. Consequently, $r(\delta)$, the Bayes risk of δ is given by

$$r(\delta) = \sum_{i=1}^{m} \pi_i \cdot R(i, \delta) \tag{2.5.2}$$

where π_i is the prior probability for class $y = i$. According to the Bayesian decision theory, the decision rule δ^* that satisfies

$$\delta^* = \arg\min_{\delta} r(\delta) \tag{2.5.3}$$

is the Bayes rule and the corresponding $r(\delta^*)$ is the Bayes risk.

It is important to note that the Bayesian optimality of the decision rule described above only applies when the gene set used for classification has been determined. Otherwise, the development of the Bayes rule requires the joint probability distribution of genes and classes. In fact, no theory is directly related to the choice of gene set for classification. Also, for a chosen gene set, the true probabilities p_{ij} and π_i are unknown in nearly any real problem. Although the Bayes rule based on empirical estimates (e.g., sample frequencies) of these probabilities is no longer guaranteed to be optimal, it has been widely considered in practice. Therefore, (2.5.1) and (2.5.2) can be used to find the prediction risk associated with δ if p_{ij} and π_i are estimated using sample frequencies $\hat{p}_{ij}$ and $\hat{\pi}_i$. As a result, different δs that correspond to the same gene set $\mathcal{S}$ can be compared to achieve the best decision, and the minimum risk resulted can be used as the score of $\mathcal{S}$. Moreover, the scores of different gene sets can be compared and searching among all possible gene sets can yield the sets with the globally minimum score – that is the basic idea of TSS.

In most problems discussed in this thesis, there is no indication of the type of loss functions and prior distributions, so the 0-1 loss function and the equal class priors are often assumed. Under 0-1 loss,

$$l_{ij} = \begin{cases} 0 \text{ if } i = j, \\ \\ 1 \text{ otherwise.} \end{cases}$$

In this case, $R(i, \delta)$ becomes

$$R(i,\delta) = \sum_{j=1}^{m} l_{ij} \cdot p(\delta = j | y = i) = \sum_{j \neq i} p(\delta = j | y = i) = 1 - p(\delta = i | y = i),$$

and $r(\delta)$ is

$$r(\delta) = \sum_{i=1}^{m} \pi_i R(i, \delta) = 1 - \frac{1}{m} \sum_{i=1}^{m} p(\delta = i | y = i).$$

Then minimizing $r(\delta)$ is equivalent to

$$\max_{\delta} \sum_{i=1}^{m} p(\delta = i | y = i) \tag{2.5.4}$$

As explained before, in most cases, only empirical solution is available for (2.5.4). Under this circumstance, the formula (2.2.3) is used to heuristically find the best decision for a given gene set. Then, all possible gene sets are sought for the globally optimal decision, which turns out to be the top scoring sets.

In addition, different loss functions or class priors can also be considered given the equation (2.5.2). For example, the empirical class prior π_i can be equal to n_i/N, where n_i is the sample size of class i and N is the total sample size, but this specification requires a good knowledge of the class distribution.

2.5.2 The Acceleration Algorithm

The acceleration algorithm described here generalizes the pruning algorithm introduced by [83] for the TSP classifier to the multiclass case. Similar to the binary TSP method, an important step for the multiclass TSS approach is to search for top scoring gene sets. Once the search process is completed, the decision rule can be immediately derived. However, given the large number of genes for microarray data, the search process is often computationally expensive. The greedy search algorithm has been introduced previsouly for searching gene sets with high scores. It is significantly more efficient than the exhaustive search, but may not be fast when combined with schemes such as cross-validation. Therefore, the acceleration algorithm here aims to expedite the search process

in the cross-validation loop.

When the greedy search algorithm is employed in the search process, only top scoring sets are kept in the final step. Therefore, gene sets that can not possibly achieve the top score can be excluded in the search process, which typically requires a "complete" comparison among all possible gene sets. In a cross-validation loop, one such comparison is needed for each iteration. However, the acceleration algorithm can produce a small list of gene sets so that only a comparison among these gene sets is sufficient to find top scoring sets. Let $r_g(n)$ denote the score obtained for a given gene set g when a subset of n samples is left out from N training samples in the cross-validation. The lower bound $L_g(n)$ and the upper bound $U_g(n)$ are defined as

$$L_g(n) \leq \min\{r_g(n) : \text{any size } n \text{ subset}\}$$

$$U_g(n) \geq \max\{r_g(n) : \text{any size } n \text{ subset}\}.$$

Now suppose the lower and upper bounds are obtained for all possible gene sets $\{g_i, i = 1, 2, ...\}$. Rank all lower bounds from largest to smallest and set the largest lower bound to L. Without loss of generality, assume $L = L_{g_1}(n)$. Then the following claim holds:

Claim: If $U_{g_i}(n) < L$ then the gene set g_i can not be a top scoring set on $N - n$ samples for any size n subset.

Proof: According to the definition of $U_{g_i}(n)$, $r_{g_i}(n) \leq U_{g_i}(n)$. If $r_g(n) \leq U_{g_i}(n) < L$, the following inequalities satisfy for any size n subset

$$r_{g_i}(n) \leq U_{g_i}(n) < L \leq r_{g_1}(n).$$

Therefore, there is at least one gene set g_1 scored higher than g_i regardless of the choice of the size n subset. This claim follows immediately.

Table 2.18: Description of the acceleration algorithm.

Acceleration Algorithm	
Input:	N training samples, gene set collection $\mathcal{G} = \{g_1, g_2, ...\}$
Output:	The reduced gene set list Ω.
1.	For each gene set g_i, compute the lower bound $L_{g_i}(n)$ and the upper bound $U_{g_i}(n)$ under all possible situations that n training samples are left out.
2.	Rank all $L_{g_i}(n)$ in descending order and take $L = \max\{L_{g_i}(n)\}$
3.	Generate the list Ω consisting of all g_i for which $U_{g_i}(n) \geq L$

The reduced list Ω typically contains only a few gene sets. The identification of top scoring sets from Ω is extremely fast. The significant improvement in efficiency is hence achieved by repeatedly using Ω in each iteration of the cross-validation. The lower and upper bound for a given gene set can be obtained by calculating all possible scores when any size n subset is left out. Unless a large n and a large number of classes are considered simultaneously, this process is also efficient.

In practice, $\mathcal{G}$ in Table 2.18 can be any gene set collection considered in the search process. For example, in the greedy search process, there are a number of sub-classification problems generated from the original problem. In each of these sub-problems, only top scoring sets are stored. Therefore, the acceleration algorithm can be applied in each step of the greedy search to yield a reduced list of gene sets that can possibly be identified as top scoring sets in the cross-validation. As a result, the greedy search process only needs to be applied one time on the training set.

The acceleration algorithm here can be immediately extended for the k-TSS classifier that uses top k scoring sets as the decision rule. In this situation, only step 2 in Table 2.18 needs to be changed so that L is set to the k-th largest lower bound. This is because any gene set whose upper bound is less than L clearly can not be one of the top k scoring sets during the cross-validation. As a result, the search process for top k scoring sets can also be quite efficient.

Chapter 3

Pathway-based Classification

3.1 Introduction

High-throughput microarray data pose challenges for most types of statistical analysis. For example, traditional hypothesis tests such as the t-test and the ANOVA F-test for selecting differentially expressed genes need to be carefully corrected for the multiple testing problem. Importantly, any statistical learning model proposed on the data has a high risk of over-fitting due to the "small n, large p" situation. Therefore, dimension reduction seems to become a necessity in microarray analysis. In early studies, univariate hypothesis tests were thought to effectively reduce the lists of genes, but different statistical procedures often produced non-identical lists, many of which had little or no overlap. As a result, there have always been concerns and discussions about the appropriateness of these procedures.

As it is well recognized that many functional related genes are typically involved in the mechanism of complex diseases such as cancers, one popular approach of gene expression analysis is to investigate these naturally defined sets of genes rather than all the genes

at once. Pathway-based classification using expression profiles has been shown in recent studies [32, 50] to provide results that are more biologically meaningful. In particular, these pre-defined groups of genes provide a natural and efficient dimension reduction. This chapter demonstrates the ability of the Top Scoring Set (TSS) approach introduced in Chapter 2 to integrate biological information from traditional pathway analysis for microarray-based classification.

Pathway Databases

There are three main sources of pathway and functional information, which can be either generic or species-specific. These might describe metabolic and cellular processes, and genetic networks. The Gene Ontology project (GO) [41] has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions. These three groups consist of genes with similar functions where genes inside each group are correlated in a hierarchical structure. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] is a series of databases developed by both the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo. The "Pathway" section of KEGG provides a plethora of searchable pathways for a diversity of organisms with the focus on metabolic pathways. GenMAPP [20] provides an image of a pathway that is annotated with accession numbers. Many gene ontology classifications are also available from GenMAPP. In addition to the databases introduced here, there are many other frequently used pathway databases such as BioCarta (www.biocarta.com) and Molecular Interaction Map (discover.nci.nih.gov/mim/index.jsp).

Enrichment Analysis

Enrichment analysis refers to the process of identifying groups of genes (rather than individual genes) that are significantly related to the disease states under study in terms of expression levels. Gene groups are typically pre-defined to share common biological function, chromosomal location, or regulation. A common case is to consider pathway information through various kinds of functional annotations of genes. Compared to the conventional identification of differentially expressed genes, pathway-level analysis is more robust to inaccuracies of specific gene expression estimates and provides a more expansive view of the underlying processes.

There are a number of statistical methods proposed for identifying pathways or functional annotations that are significantly associated with phenotypes under study based on expression values. In the most common approach, genes are first ordered according to their evidence for differential expression, by one of many univariate statistical tests. Then they are examined against each of the gene sets defined by pathways, to determine whether any set is over-represented (i.e., contains more top genes than other sets) among the whole list of genes. To examine the evidence of association in this case, Fisher's exact test based on the hypergeometric distribution or its large-sample approximation χ^2 test is a routine. However, as mentioned in [84], there are at least three shortcomings for this approach, and a better solution for enrichment analysis is in need.

A more successfully and widely recognized method is Gene Set Enrichment Analysis (GSEA) proposed by Subramanian et al. [82]. The method was originally designed for two-class problems. It works in following steps for a collection of gene sets $S_1, S_2, .., S_n$:

1. Compute the t-statistic (z-score) z_j (comparing across the two classes) for each gene j of all N genes in the data.

2. Generate a summary statistic for each gene set S_k. In GSEA, this statistic is defined as the enrichment score that is essentially a signed version of the Kolmogorov-Smirnov statistic between the values $\{z_j, j \in S_k\}$ and their complement $\{z_j, j \notin S_k\}$. The sign is positive if $j \in S_k$ and negative if $j \notin S_k$.

3. Permute the sample labels and recompute the statistic on each the permuted dataset, which generates a null distribution of the enrichment score.

4. Assess the statistical significance (P value, false discovery rate, etc.) of the observed score and significant gene sets are chosen.

GSEA provides a statistical framework for comparing different pathways, and it has been extended in many subsequent studies. For instance, Zahn et al. [93] considered a Van der Waerden statistic in place of the Kolmogorov-Smirnov statistic, and bootstrap sampling of the arrays instead of a permutation distribution. Significance Analysis of Function and Expression (SAFE) proposed by Barry et al. [4] extends GSEA to cover multiclass, continuous and survival phenotypes, and gives two more options for the test statistic: the Wilcoxon rank sum and the hypergeometric statistic that uses the Fisher's exact or its large-sample approximation χ^2 test as mentioned above. Efron and Tibshirani [25] introduced the Gene Set Analysis (GSA) method that uses a new statistic called "maxmean", and developed a restandardization step for producing more accurate enrichment scores.

Microarray Classification

The successful identification of differentially regulated pathways associated with phenotypes of interest through enrichment analysis advances the development of molecular

classification of diseases. Numerous studies have relied on analysis of individual genes to find microarray-based molecular markers for discrimination of disease states (e.g., subtypes and therapy responses) in experimental and clinical settings. However, such procedures in complex diseases are often limited by factors such as cellular heterogeneity within a tissue sample and genetic heterogeneity across patients. As a result, the effectiveness and reproducibility of many early results have been challenged in follow-up studies. Moreover, lack of coherence in biological interpretation of these results also prevent them from being translated into clinical applications. In view of these limitations, a growing trend is to integrate pathway information into the biomarker identification process, permitting disease classification based on the activity of pathways rather than simply on the expression levels of individual genes. This section gives a short introduction to some of the approaches developed for this purpose.

First of all, based on the enrichment methods described in the last section that lead to detection of statistically significant pathways, genes in those pathways can be considered for a careful investigation for their association with the disease under study. There is much evidence in recent studies of correct identification of mechanism-related genes based on pathway analysis [26, 52]. Second, compared to traditional biomarkers, pathway-based gene signatures tend to have a high level of concordance, i.e., consistency in terms of the gene members contained or agreement of analysis results using these signatures. For example, Fan et al. [28] examined the predictions derived from a number of prognostic gene signatures for individual samples related to breast cancer, and found that most models had high rates of concordance in their outcome predictions. As a matter of fact, nearly all these gene signatures have incorporated some of the main signaling pathways of breast cancer. Therefore, pathway analysis can play an important role in disease

classification. In particular, diagnostic and prognostic gene signatures can be sought directly from statistically significant or mechanism-related pathways for groups of genes collectively exhibiting consistent expression patterns that allow for distinction among phenotypes.

In addition to producing gene signatures directly, pathways are more involved in a variety of statistical learning models developed for disease classification. In this situation, each pathway is often treated as a functional module where a summary measure is computed to capture the overall activity level of the module. To be precise, for a certain patient sample, each module (pathway) activity level is a scalar generated by applying some mathematical transformation of the raw expression values of genes contained in that module. As a result, the original high-throughput data is transformed into low dimension functional expression profiles, on which many conventional classification models are built. For instance, Guo et al. [36] considered decision trees on functional expression data where module activities were computed by simply taking the mean or median of expression values of module genes. Lee et al. [52] proposed a logistic regression model in which each activity level was inferred as an averaged z-score derived from the expression of its individual key genes that generated most discriminative activities between two phenotypes. A pathway activity matrix was constructed based on activity levels to train a classifier. Also, Su et al. [81] inferred a particular pathway activity by first computing the log-likelihood ratio between different disease phenotypes based on the expression level of each gene, and combining the log-likelihood ratios of the genes contained in that pathway. Then, they considered a regular logistic regression or a linear discriminant analysis model for disease classification. In a recent study by Kim et al. [50], the support vector machine (SVM) technique was used on 2-level hierarchical data with a basic level of gene

expression values and an advanced level of pathway activities. In this case, each pathway activity level was a linear combination of expression values of its member genes and the weights used were generated by applying a SVM with the whole set of genes.

In summary, pathway-based classification can be performed in two major ways. Within disease-related or statistically significant pathways, small sets of key genes can be identified in association with different phenotypes. The expression patterns of these genes can lead to accurate and biologically meaningful disease classification. At the same time, numerical patterns of a certain pathway can be inferred and summarized into an activation level using the expression values of the constituent genes, which generates a low dimensional dataset with features being the pathway activities instead of gene expression values. As a result, conventional statistical learning methods become immediately applicable. In this chapter, the Top Scoring Set method introduced in Chapter 2 is considered to integrate biological information from pathway analysis, and a new pathway-based classification approach is proposed later.

Unsupervised Learning

While this thesis is primarily focused on supervised learning problems, part of this chapter is dedicated to unsupervised learning tasks. For studying gene expression microarrays, unsupervised techniques are constantly used for identifying unexpected but biologically interesting patterns in the data. In particular, they are of great importance to identify clusters in samples when no prior classification of samples is available. In this chapter, a rank-based unsupervised approach is introduced that is based on hierarchical cluster analysis, and aims to address the robustness of current clustering techniques as applied to gene expression microarrays. The method uses the Kendall's rank coefficient [48]

as the distance metric, which is consistent with the rank-based transformation used in other methods in this thesis, and provides a complementary tool to the family of relative expression analysis methods especially when no prior classes of samples are available or reliable. In addition, the method is suitable to combine with rank-based classification methods such as TSP or TSS. This unsupervised approach is described in detail in the "Methods" section of this chapter. The "Results" section provides a validation study of using this method to investigate the molecular subtypes of breast cancer patients.

3.2 Methods

3.2.1 Pathway-based Top Scoring Set

The TSS method introduced in Chapter 2 provides an effective and transparent classification model based on gene expression levels. It explores the discriminative power of the relative expression comparison in small-sized gene sets where subtle but consistent expression changes of a few genes across all samples lead to class separation. However, TSS is somewhat limited by the fact that it is not computationally feasible to carry out an exhaustive search and the global optimality requirement needs to be relaxed. One way to address this problem would be to pre-select a small number of genes based on a univariate multiclass criterion such as one-way ANOVA. Another approach proposed in Chapter 2 is the greedy search algorithm. However, all these methods are exclusively based on the expression patterns of genes, and the reliability of results derived in small-sample settings could be questionable due to the inherent measurement noise in high-throughput data and the heterogeneity across samples and patients. Motivated by these considerations, this section introduces a new classification procedure that combines TSS with biological

knowledge achieved from pathway analysis.

Pathways can provide biological interpretations while restrict attention to small-sized (tens to hundreds) gene groups. For the TSS approach, each such group of genes offers a unique opportunity to search for top scoring sets and to uncover subtle changes of co-expression pattern in different phenotypes. In fact, the relative comparison idea used by TSS could make more sense based on pathways where dependency structures among genes are known to exist. Also, top scoring sets obtained by searching within pathways could provide immediate biological insights into the underlying disease-related mechanisms.

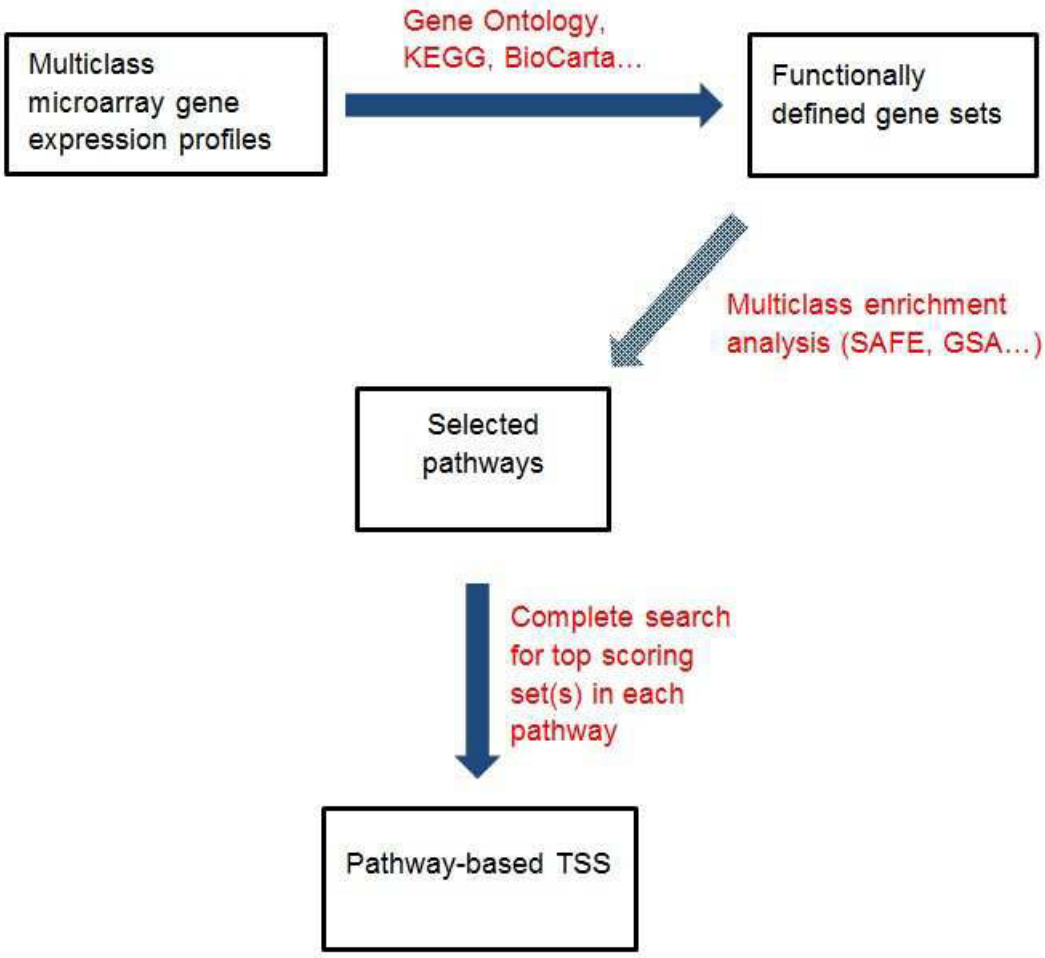

Figure 3.1: Scheme of the pathway-based Top Scoring Set method.

The pathway-based TSS method is described in Figure 3.1. The procedure begins

with a microarray gene expression data set and finds the functional modules defined by a particular pathway database such as Gene Ontology or KEGG using the annotations of all genes. Then, an *optional* step is to apply some enrichment technique to rank all pathways according to their discriminative ability among different phenotypes, and pathways that are identified as statistically significant are chosen. Next a complete search within each of these pathways is performed to obtain high-scoring gene sets. Finally, the sets from different pathways are compared to yield top scoring sets as the ultimate classification rule.

The restriction to pathways greatly improves the efficiency of the search process. It breaks down the large feature space into small groups of genes where relative expression comparison is appropriate to consider within each group. These gene groups are pre-defined using biological knowledge so that they can be more objective and meaningful than those obtained by statistical hypothesis tests. Importantly, a pathway consists of genes that connect and interact with each other, which can make the exploration of relative expression analysis more promising than traditional methods.

The enrichment method used above requires multiclass analysis. In binary problems, the celebrated GSEA method is the most common choice. In multiclass settings, there are also a number of approaches that follow the similar procedures of GSEA and they differ from each other in several aspects. First, the local statistic calculated for each gene in the data can be different. The most common choice for this is the ANOVA F-statistic, and the Kruskal-Wallis test statistic can also be an alternative. Second, the summary statistic for each gene set can be different. Besides the Kolmogorov-Smirnov statistic considered in GSEA, other metrics such as Wilcoxon rank sum (used in the SAFE method) can also be reasonable options. Finally, the methods for assessing statistical significance of

summary statistics can be different. The null distribution can be generated by permuting class labels or creating bootstrap datasets.

In fact, as indicated in Figure 3.1, enrichment analysis is only optional since the statistical significance of a pathway is not directly associated with the scores of gene sets formed. In cases where there are a large number of pathways associated with genes studied, enrichment methods can be introduced to filter out undesirable pathways that contain noisy or no signals. However, the criterion of choosing *significant* pathways can vary among problems and methods used, which may lead to totally different sets of pathways selected. Therefore, for databases like KEGG where a relatively small number of pathways (<300) is often identified, the enrichment analysis step is not necessary.

For some problems studied in this chapter, pathway-based TSS is applied in a similar fashion to the k-TSP method introduced by Tan et al. [83]. To be precise, gene sets with top k scores from the pathway search are combined to form the final classification rule. For complex multicategorical diseases, the decision from a single gene set is unlikely to be robust and effective, and it seems necessary to make ensemble decisions from a group of high-scoring gene sets. In practice, a choice of the maximum number of gene sets considered in the decision rule (denoted by $k_{\max}$) is often specified. The optimal $k(k \leq k_{\max})$ is typically determined by using leave-one-out cross-validation (LOOCV) on the training set, a common procedure for small-sized data [33,56]. However, LOOCV typically requires extensive computations, especially for estimating the accuracy of a classification method with parameters (e.g., the choice of k here). Therefore, the acceleration algorithm described in Section 2.4.2 is also used to improve the efficiency of training such a pathway-based k-TSS classifier. As discussed in Section 2.4.2, the acceleration algorithm can be used for any collection of gene sets and can be immediately extended to select candidate

gene sets for a k-TSS classifier. In particular, the acceleration algorithm used for the pathway-based k-TSS classifier is described in Table 3.1.

Table 3.1: The acceleration algorithm for the pathway-based k-TSS classifier.

Acceleration Algorithm	
Input:	N training samples from m classes, a pre-defined pathway $\mathcal{S}$ and $k_{\max}$
Output:	The reduced gene set list Ω.
1.	Generate the collection of gene sets (size m) by considering all possible combinations of genes from $\mathcal{S}$. Denote the collection as $\mathcal{G} = \{g_1, g_2, ...\}$.
2.	For each gene set $g_i \in \mathcal{G}$, compute the lower bound $L_{g_i}(n)$ and the upper bound $U_{g_i}(n)$ under all possible situations that n training samples are left out. Set $n = 1$ for leave-one-out cross-validation.
2.	Rank all $L_{g_i}(n)$ in descending order and set L as the $k_{\max}$-th largest value of all $L_{g_i}(n), i = 1, 2, ...$
3.	Generate the list Ω consisting of all g_i for which $U_{g_i}(n) \geq L$

Notations are the same as defined in Section 2.4.2. Note that multiple pathways can also be considered by adding all possible gene sets (e.g., combinations of genes) from these pathways to the collection of gene sets $\mathcal{G}$. Also, the resulting gene set list Ω is used in each iteration of the LOOCV process where top scoring sets are found among all gene sets in Ω.

3.2.2 Rank-based Clustering

Hierarchical clustering is one of the most popular techniques for microarray analysis. It is often used to arrange samples according to similarity in patterns of expression. The two basic building blocks in hierarchical clustering are the *distance matrix* and the *linkage method*. The distance matrix consists of pairwise distances between samples to be clustered. A number of sensible ways of defining the distance metric between two samples can be used when samples have numerical-valued vectors associated with them, including the Euclidean distance and the Pearson's correlation coefficient. The linkage method determines the distance between two groups of samples as a function of pairwise distances. For example, the average linkage method defines the distance between two clusters A and B as

$$\frac{1}{|A||B|}\sum_{i\in A}\sum_{j\in B} d(i,j)$$

i.e., the average of distances between any possible pair from A and B respectively. Based on these two elements, hierarchical clustering uses the agglomerative algorithm to form clusters in a bottom-up manner where two clusters with the smallest distance at each step is merged. In general, hierarchical clustering has the distinct advantage that the samples are not needed once the distance matrix is constructed.

In this section, a rank-based clustering approach is introduced for gene expression analysis. In particular, the distance matrix associated with samples is calculated based on relative ranks of gene expression levels and it is used in the hierarchical clustering process. In the context of microarray analysis, value-based distance metrics such as the Euclidean distance or the Pearson's correlation coefficient rely entirely on the actual expression values of the genes involved, which could be sensitive to subtle variations in data preprocessing and the proportion of samples from each class. Hence, the goal here is

to improve the robustness of hierarchical clustering by considering a rank-based distance metric.

The rank correlation coefficient developed by Kendall [48] is an excellent choice for the distance metric. This distance, sometimes referred to as the Kendall tau distance, counts the number of pairwise disagreements between two ranking lists. The smaller the distance, the more pairs of values remain the same relative ordering for two lists, which indicates the two lists are more similar. The Kendall tau distance has already been investigated by Afsari [1] for its use as a rank-based classifier for distinguishing cancer phenotypes (See Section 2.5 in [1]). Here its potential for unsupervised clustering is explored.

Now suppose an expression matrix $\{x_j^{(i)}\}$ is given for $i = 1, 2, ..., N$ samples and $j = 1, 2, ..., p$ genes. Samples are presumably from a number of phenotypes (e.g., tumor subtypes), and genes are chosen so that they can be potentially related to phenotypes of interest. The Kendall tau distance d between any pair of samples, say i_1 and i_2 is defined as

$$d(i_1, i_2) = \frac{1}{\binom{p}{2}} \sum_{\substack{j_1, j_2 \in \{1,2,...,p\} \\ j_1 \neq j_2}} d^*_{j_1, j_2}(i_1, i_2)$$

where

$$d^*_{j_1, j_2}(i_1, i_2) = \begin{cases} 1 & \text{if } \arg\max_{j_1, j_2}\{x_{j_1}^{(i_1)}, x_{j_2}^{(i_1)}\} \neq \arg\max_{j_1, j_2}\{x_{j_1}^{(i_2)}, x_{j_2}^{(i_2)}\} \\ 0 & \text{otherwise} \end{cases}$$

Basically, $d(i_1, i_2)$ measures the proportion of mismatches among a group of genes where a mismatch is counted if a certain gene pair of sample i_1 has a different maximum expressed gene (i.e., gene whose expression level is the highest) from that of sample i_2. In fact, this distance could also be defined by considering mismatches of gene sets containing three, four or any reasonable number of genes, but the computational complexity is expected

to increase significantly.

By definition, if two samples have a small value of $d(i_1, i_2)$, most of their two-gene sets tend to have the same gene as the maximum expressed gene. As a result, they are expected to have the same relative ordering for most two-gene sets. Also, it is quite possible that the relative ordering of gene sets of larger sizes (>2) are also preserved. In addition, a small $d(i_1, i_2)$ also indicates a high probability that there exists a two-gene set with the same maximum expressed gene across samples i_1 and i_2, which is related to the idea of the TSP approach. When the mismatch is considered for gene sets of size greater than two, the distance definition is related to the TSS approach.

A distance matrix D for N samples (and p genes) can be computed based on the Kendall tau distance. Supervised classification rules can then be constructed based on D. For instance, Afsari [1] considered the average swap distance as a classification rule that is equivalent to the following rule for a sample l in terms of the Kendall tau distance

$$\underset{c=1,2,\ldots}{\arg\min} \quad \frac{1}{n_c} \sum_{i \in \text{class } c} d(i, l)$$

where n_c is the number of samples in class c. Basically, the rule computes the average Kendall tau distance between sample l and those in each class, and the prediction is the class with the minimum average distance.

For unsupervised problems, the distance matrix D can be used in a similar way as for supervised learning. The main assumption is that samples from different phenotypes are expected to have larger distances than those from the same phenotype, which is intended to indicate different genetic perturbations (e.g., activation of different pathways) existing in the underlying mechanism of a certain disease. Based on the matrix D, hierarchical cluster analysis can be immediately applied with a linkage method, and nothing else is needed for clustering.

The Kendall tau distance is a well-defined metric satisfying the following properties for any samples i, j and k:

- $d(i,j) \geq 0$ (non-negativity)
- $d(i,j) = 0 \Leftrightarrow i = j$ (identity)
- $d(i,j) = d(j,i)$ (symmetry)
- $d(i,k) \leq d(i,j) + d(j,k)$ (triangle inequality)

In addition, this distance can also be derived from the Hamming distance [39] that has been frequently used in information theory.

In the following sections of this chapter, the pathway-based TSS classifier and the rank-based clustering approach are validated in two settings. The first one is to distinguish subtypes based on an extremely large cohort study related to leukemia cancer. The second one relates to the building of a subtype predictor for the prognosis of breast cancer patients.

3.3 Results

3.3.1 Classification of Leukemia Subtypes

The study of leukemia cancer via gene expression analysis is the setting for one of the earliest attempts in microarray-based cancer classification. In the prominent work by Golub et al. [35], human acute leukemias were chosen as a test case where a microarray-driven classification model was built based on 38 patient samples. The motivation came from evidence showing that acute leukemias can be classified into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute

myeloid leukemia, AML), and no single test could efficiently establish the diagnosis. The classification model derived included 50 informative genes and was tested on an independent collection of 34 leukemia samples. 30 out of 34 samples were correctly predicted for the ALL-AML distinction. Motivated by this work, the usefulness of microarray-based gene expression classification has then been widely recognized and explored.

In this section, the first attempt was to test the pathway-based TSS (P-TSS) approach on the same leukemia data from the Golub's study using the same protocol for training (38 samples) and testing (34 samples). Pathway information was collected from KEGG, and the Bioconductor package [34] in R was used to find 226 KEGG pathways on the data. To assess the statistical significance of these pathways, the GSA approach [25] was used to perform enrichment analysis. Table 3.2 lists all differentially expressed pathways identified by GSA when the false discovery rate (FDR) threshold is set to 0.2 and the size (i.e., the number of genes) of a pathway is limited to between 15 and 500 by default. A TSS classifier was built by searching within each of these five pathways, and it finally contained five top scoring sets with the score of 2.94. The classification result on the independent testing set was compared to that of benchmark classifiers in Table 3.3.

P-TSS achieves the highest accuracy, and it used the smallest number of genes. The top scoring sets in the final decision rule come from two pathways, namely, primary immunodeficiency and hematopoietic cell lineage, both of which are statistically significant with zero P value using GSA, and they also have interesting biological relevance. The former one relates to disruption of the cellular immunity observed in patients with defects in T cells or both T and B cells, and the latter one is linked to the manner in which blood-cell development progresses from a hematopoietic stem cell.

While the result above demonstrates the effectiveness of P-TSS, due to small sample

Table 3.2: Significant KEGG pathways identified on leukemia samples.

Gene set	Size	P value	FDR
Primary immunodeficiency	32	< 0.001	< 0.001
T cell receptor signaling pathway	96	< 0.001	< 0.001
B cell receptor signaling pathway	68	< 0.001	< 0.001
Hematopoietic cell lineage	104	< 0.001	< 0.001
Rheumatoid arthritis	88	0.005	0.178

Table 3.3: Comparison of classification methods on leukemia samples.

Method	Test error	No. of genes
Golub et al.	4/34	50
kNN	4/34	7129
NB	4/34	7129
RF	4/34	1273
PAM	1/34	47
l-SVM	5/34	7129
P-TSS	1/34	10

there is reason to anticipate a lack of robustness and reproducibility. To address this issue, P-TSS was also tested on the microarray data from the MILE (Microarray Innovations in Leukemia) study [37] described in Section 2.3.1, with the purpose of establishing an accurate and robust class predictor. The large sample size and the strict data quality

criteria utilized for MILE provides an excellent opportunity for validation. In particular, two classification problems (Table 3.4) were generated from the MILE data: one to distinguish three major lineages (B-ALL, T-ALL and AML) in acute leukemia samples, and the other to differentiate CLL (chronic lymphocytic leukemia), CML (chronic myelogenous leukemia) and MDS (myelodysplastic syndromes). For these two large sample problems, the k-TSS approach that makes ensemble decision based on the top k scoring sets were used. The value of k with an upper bound of $k_{\max} = 50$ was determined using leave-one-out cross-validation (LOOCV) on the training set.

Table 3.4: Samples of leukemia subtypes used for classification.

Class	No. of samples	
	Training	Test
B-ALL	576	357
T-ALL	174	79
AML	542	257
CLL	448	237
CML	76	43
MDS	206	121

For the first problem, the training began with a pathway analysis on all 1,457 (non-missing) genes that are common for the training and test set. Information was collected from KEGG database to yield 56 pathways, each of which was required to contain at least 10 members in order to obtain robust signals. Next, the k-TSS approach was applied by searching within 56 pathways. The acceleration algorithm was used to generate a list of

57 candidate gene sets, and the LOOCV accuracy for each k-TSS classifier ($k \leq k_{\max}$) was estimated based on the list.

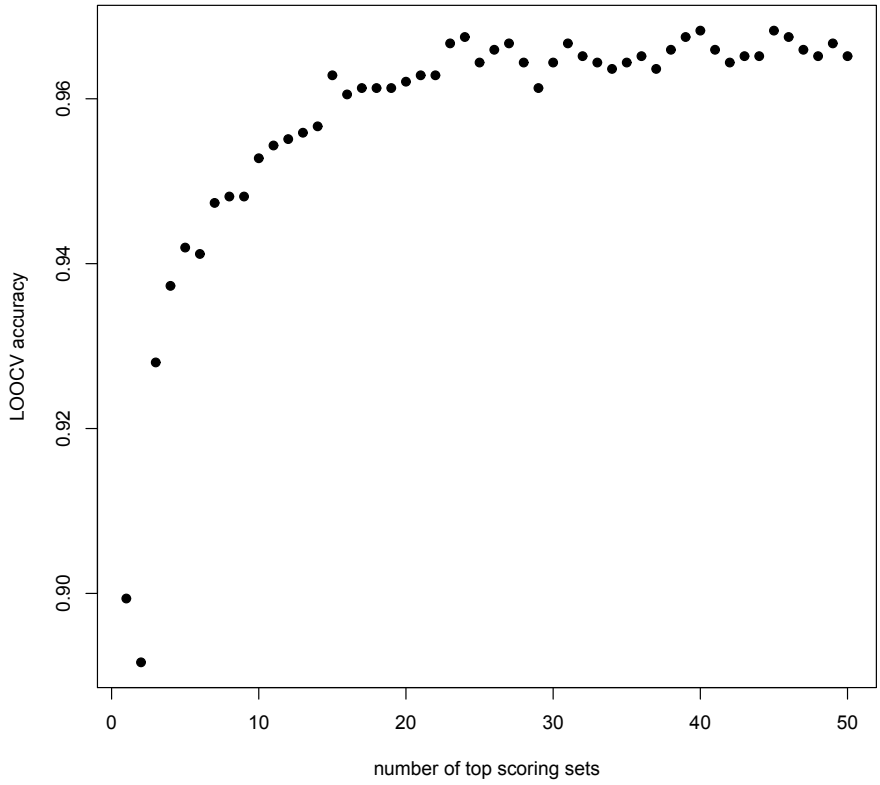

Figure 3.2: LOOCV accuracies of all k-TSS classifiers on acute leukemia samples.

Figure 3.2 compares the LOOCV accuracy of k-TSS classifiers on 1,292 training samples. The optimal k was found to be 40 with the maximum LOOCV accuracy of 96.83%. Therefore, a TSS classifier was built on the top 40 scoring sets in the training process. Furthermore, an attempt was made to find the pathways in which these 40 gene sets were formed. Interestingly, all 40 gene sets are actually from only two out of 56 pathways: primary immunodeficiency and hematopoietic cell lineage. This result is remarkably consistent with that obtained on the leukemia data from the Golub's study where all top scoring sets come from the same two pathways. Notice that this consistency is achieved

on data sets from two different platforms (Affymetrix HuGeneFL and HG-U133 Plus 2.0), which seems to indicate the biological importance of these pathways in the distinction of three acute leukemia subtypes.

The second problem is to distinguish among two chronic leukemias (CLL, CML) and MDS samples. Since the set of genes is the same as for acute leukemia samples, there are still 56 KEGG associated pathways. In the training process, the acceleration algorithm produced a list of 70 candidate gene sets and the LOOCV accuracies for all 50 k-TSS classifiers are displayed in Figure 3.3. The optimal model turns out to be the 7-TSS classifier with the highest accuracy of 97.26%. The top seven scoring sets are identified from three pathways listed in Table 3.5.

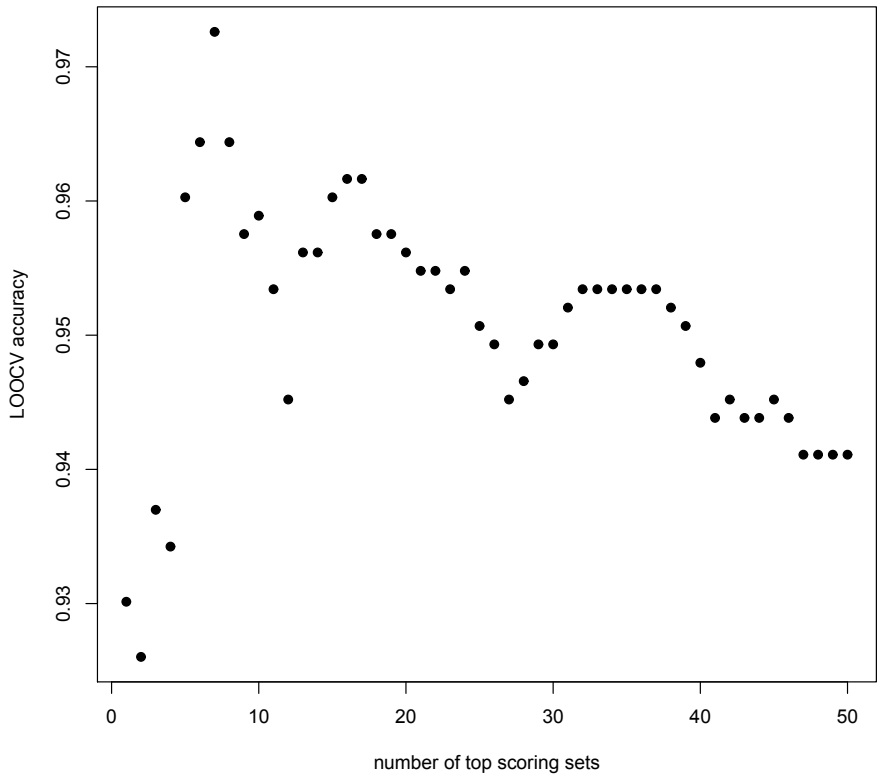

Figure 3.3: LOOCV accuracies of all k-TSS classifiers on chronic leukemias and myelodysplastic syndromes samples .

After training, the two k-TSS classifiers were tested on samples from stage II of the

Table 3.5: Three KEGG pathways identified for chronic leukemias and myelodysplastic syndromes samples.

Pathway	Size	No. of TSS	Averaged score
Metabolic pathways	59	4	2.71
Pathways in cancer	47	1	2.67
Hematopoietic cell lineage	102	2	2.66

MILE study. In particular, two advanced machine learning techniques were selected as benchmark methods: linear kernel support vector machines (l-SVM) and Prediction Analysis of Microarray (PAM). The same implementation of these two methods as in the "Classification of Human Cancer Microarray Data" section were used. Both benchmark classifiers were optimized on the same training data through cross-validation, and the fold was restricted to 10 to save computation time.

Validation results are provided in Table 3.6 and 3.7. For the first problem, P-TSS achieves high accuracies, and the rate for B-ALL, T-ALL and AML is 98.0% and 93.7% and 95.3% respectively. High accuracies are also observed for l-SVM and PAM. l-SVM and PAM seems to outperform P-TSS in class AML, but are less effective in B-ALL. P-TSS generally yields comparable accuracies to benchmark approaches, and it only uses 25 genes for classification while PAM uses 845 and l-SVM uses all 1457 genes. Moreover, the decision rule for P-TSS is transparent and may be interpretable, which could have some advantage over the complex decision boundaries generated by l-SVM and PAM.

Also, for the second problem, P-TSS obtains accuracies of 98.7%, 93% and 95.9% for CLL, CML and MDS respectively. l-SVM turns out to be the best classifier on this data with highest accuracies in all three classes. PAM also yields high accuracies in CLL and CML, but for MDS accuracy is significantly lower than the other two methods. In addition, l-SVM and PAM both uses all 1457 genes while only 21 genes are involved in

Table 3.6: Confusion matrices on acute leukemias.

P-TSS

Actual ↓ / Predicted →	B-ALL	T-ALL	AML
B-ALL	350	0	7
T-ALL	2	74	3
AML	3	9	245

l-SVM

Actual ↓ / Predicted →	B-ALL	T-ALL	AML
B-ALL	347	2	8
T-ALL	2	74	3
AML	1	3	253

PAM

Actual ↓ / Predicted →	B-ALL	T-ALL	AML
B-ALL	339	5	13
T-ALL	2	75	2
AML	1	7	249

the P-TSS classifier, and P-TSS again produces very comparable results.

Summary

The pathway-based TSS approach is proposed to search for top scoring sets whose member genes are from the same pathway defined by a certain database. In this section, this approach is trained and validated on two microarray data sets related to leukemia cancer. The first one is a widely studied group of 72 leukemia patients. In this case, P-TSS used only five gene sets from two KEGG pathways for classification and outperformed all benchmark methods. The second one is from the MILE project, one of the largest

Table 3.7: Confusion matrices on chronic leukemias and myelodysplastic syndromes.

P-TSS

Actual ↓ / Predicted →	CLL	CML	MDS
CLL	234	1	2
CML	1	40	2
MDS	0	5	116

l-SVM

Actual ↓ / Predicted →	CLL	CML	MDS
CLL	236	1	0
CML	0	43	0
MDS	3	1	117

PAM

Actual ↓ / Predicted →	CLL	CML	MDS
CLL	236	0	1
CML	0	42	1
MDS	1	22	98

cohort studies of leukemia cancer. The top k scoring sets obtained by searching within all available KEGG pathways were combined for an ensemble classification where k was optimized on the training set by LOOCV. The effectiveness and robustness of P-TSS were demonstrated on two data sets generated from the MILE study.

The results above show that the P-TSS classifier is reliable for classification of leukemia subtypes. Previously, top scoring sets have been sought using all available genes, and one possible issue is the over-fitting problem. Here, although each pathway typically contains only tens to hundreds of genes, gene sets with high scores can still be found and have been proved to be effective for predicting new samples. It is interesting that these gene

sets may not come from pathways identified as statistical significant by any enrichment method, which implies the difference between the traditional expression analysis and the relative expression analysis.

Nonetheless, there are several limitations for results obtained in this section. One limitation is the set of genes used for the MILE data. In fact, the training set from stage I contains 54,675 genes, but only 1,457 genes that are shared by the validation set are used in the training process. This fact significantly restricts the number of possible gene sets that are formed from pathways. Also, since all expression data sets considered here are related to leukemia cancer, it is desirable to have further study on other types of diseases.

3.3.2 Breast Cancer Prognosis through Subtype Prediction

Background

Microarray-based expression studies have uncovered that breast cancer is both a clinically diverse and molecularly heterogeneous disease [38]. Although most treatment decisions are still based on clinical-pathological factors such as age, tumor size, histological grade, lymph node metastasis, estrogen receptor and progesterone receptor status, etc., molecular subtypes of breast cancer defined by distinct gene expression patterns have been demonstrated by various research groups [65, 76, 77] to be associated with clinical outcome. In the last decade, several statistical models have been proposed to identify breast cancer subtypes at the molecular level where different subtypes tend to have different risks of relapse/survival or responses to chemotherapy. Although the stability of these

predictive models is still unclear and being tested, their concordance with respect to the predictions for individual patients has been demonstrated by Fan et al. [28]. In addition, a high rate of concordance has also been reported between the identified subtypes and the risk classifications generated by well-known gene signatures such as Mammaprint [86], Oncotype DX [61] and the Gene expression Grade Index (GGI) [78]. As a result, there is an increasing need for the development of independently validated methods for the identification of these subtypes.

According to the seminal work of Perou et al. [64], breast cancer tumors can be classified into at least four subclasses: luminal, basal-like, HER2-ennriched (HER2+) and normal breast-like. This classification is based on a large set of "intrinsic" genes (i.e., genes exhibiting little variance within repeated samples of the same tumor while having high variance across different tumors) and these classes have been termed "intrinsic" subtypes. Some subsequent studies confirmed the identification of these subtypes and showed that the luminal group could be further divided into two (luminal A and B) or three (luminal A, B and C) subclasses. Also, Desmedt et al. [15] studied a number of biological processes including estrogen receptor (ER) and HER2 signaling and derived a slightly different classification of breast cancer tumors consisting of three main subtypes: ER-/HER- (basal-like), HER2 (HER2+) and ER+/HER- (luminal) where luminal subclasses were combined into a single one and the normal breast-like group was left out.

The most commonly used technique for identifying molecular subtypes of breast cancer is hierarchical clustering, which allows a simultaneous analysis of many genes for detecting co-expressed groups. However, as discussed in [38], this approach requires samples in large retrospective studies and can not be employed for identifying subtypes for samples from individual patients. Therefore, several statistical models called "single sample pre-

dictors"(SSPs) have been developed to assign the subtype of a single tumor. Although these models have been based on different lists of intrinsic genes, their classification rules are essentially of the nearest centroid classifier type that computes the "distance" between a given sample and the centroid of each subtype based on expression values of intrinsic genes, and selects the "nearest" subtype. In the latest derivation of SSP, the PAM50 model employed the nearest shrunken centroid approach PAM (introduced by Tibshirani et al. [85]) and was based on only 50 genes.

Despite the considerable success achieved by SSP models, their limitations have been showed and discussed in several studies. One major issue lies in the statistical methods used for subtype identification and classification [38]. First, the clusters detected in hierarchical clustering based on expression values are not stable under subtle variations in data preprocessing as well as the choices of samples. Second, subtype classifications have been reported to depend on the list of intrinsic genes [88], and different SSPs only yielded moderately concordant classifications. Due to these limitations, some alternative subtype models were also proposed that are not based on any intrinsic gene list such as the Subtype Classification Model [38].

In this section, the ability of the rank-based clustering approach to independently identify previously defined subtypes is investigated. The goal here is not to demonstrate the superiority of this approach in terms of clustering or classification, but to validate the possible existence of "intrisic" breast cancer subtypes and to provide one of many explanations for the structural difference among these subtypes through the relative expression analysis. As discussed above, although breast cancer is commonly known to be molecularly heterogeneous, the lack of "gold standard" classifications of breast cancer subtypes largely limits the validation of various subtype classification methods/models.

The study carried out in this section adopts one of the most notable lists of intrisic genes used by the PAM50 model to explore the existence and the consistency of the previously identified subtypes. At the same time, the k-TSP classifier is proposed for the prediction of individual tumors based on the subtypes identified. Both the clustering and classification approach are compared to the PAM50 model using commonly used statistical tests.

Gene Expression Data

A collection of three gene expression data sets (VDX, MAINZ and TRANSBIG) provided in [38] were used. These data sets were assayed on the same platform and they contain 742 untreated and node negative patients with distant metastasis-free survival data available. None of these data sets were used to derive the PAM50 model, but VDX was used for training the rank-based clustering. k-TSP and PAM50 were applied on MAINZ and TRANSBIG for subtype prediction and survival analysis. Other information about these data sets is provided in Table 3.8.

Table 3.8: Three gene expression datasets of breast cancer patients.

Dataset	Platform	No. of genes	No. of samples	Source	Reference
VDX	Affymetrix HGU133A	22283	344	GSE2034/GSE5327	[87]
MAINZ	Affymetrix HGU133A	22283	200	GSE11121	[73]
TRANSBIG	Affymetrix HGU133A	22283	198	GSE7390	[16]

Subtype Discovery

The class discovery of breast cancer tumors is a major challenge to the development of molecular subtype models. Hierarchical cluster analysis is often employed to handle this

unsupervised learning task, but its biostatistical limitations and numerical instability have been well documented [67]. For the prevalent SSP models, clustering results depend on the list of intrinsic genes. Moreover, Mackay et al. [58] pointed out the lack of objectivity and interobserver reproducibility in manually identifying subtypes from dendrograms yielded by hierarchical clustering. As a result, there are no clear guidelines for describing in detail how each molecular subtype should be identified from the visual analysis of the dendrograms obtained from hierarchical cluster analysis.

Despite these difficulties, as recently discussed by Perou et al. [64], the intrinsic subtype model is an evolving classification system. PAM50 should be preferred over earlier SSP models and its predictive power as well as prognostic value is likely to be confirmed as more and more independent validation studies are carried out. Therefore, an attempt has been made for discovery of breast cancer subtypes based on the intrinsic gene list of PAM50 while using the rank-based hierarchical clustering. The rank-based approach adopts a completely different methodology to define the distance metric, which makes use of the relative ordering instead of the actual expression values of the intrinsic genes. Two samples are defined as "close" to each other if their relative orderings match for most possible gene pairs.

The list of 50 intrinsic genes used by PAM50 was obtained from the R package *genefu*. Genes were mapped using Entrez GeneID. The probe with highest variance was used when multiple probes mapped to the same GeneID. 44 out of 50 genes were found in the probe set of three gene expression data in Table 3.8. A (rank-based) distance matrix was constructed for 344 samples in VDX based on 44 genes. The average-linkage hierarchical clustering was applied and the dendrogram was displayed in Figure 3.4.

At the highest level, the samples seemed to be clustered into four large groups. The

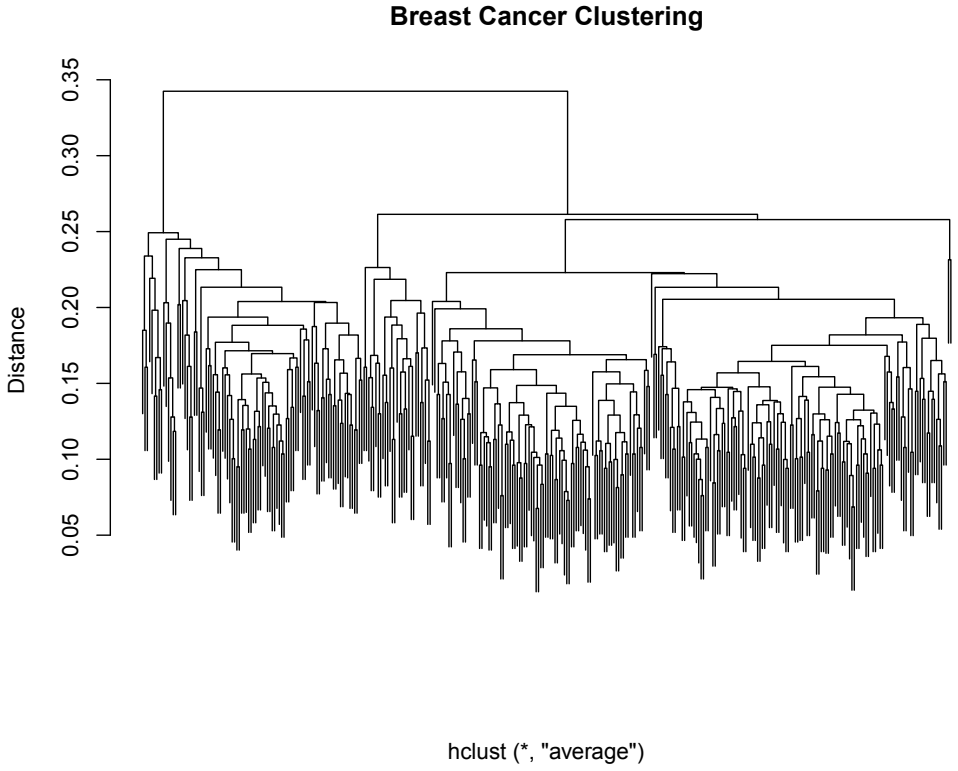

Figure 3.4: Rank-based hierarchical clustering of breast cancer tumors.

leftmost group is separated from the other three with a distance of roughly 0.35. The second group from the left is the smallest and is at least a distance 0.25 apart from the two large groups on the right. The average distance in each group is around 0.15, which indicates that 85% of possible relative orderings (in fact, that is $0.85 \cdot \binom{44}{2} \approx 804$) are matched between two samples from the same group. Previous studies have shown that there are at least four breast cancer subtypes, so it is desirable to estimate the concordance between the previously defined subtypes and the clusters found in Figure 3.4.

Figure 3.5 is colored using the five subtypes predicted by PAM50. All basal-like tumors seemed to be included in the leftmost group, together with all normal-like tumors. All expect one HER2+ tumor is in the second group (from the left). The next group contains all luminal A tumors and the last one is a mixture of luminal A and B tumors. The five

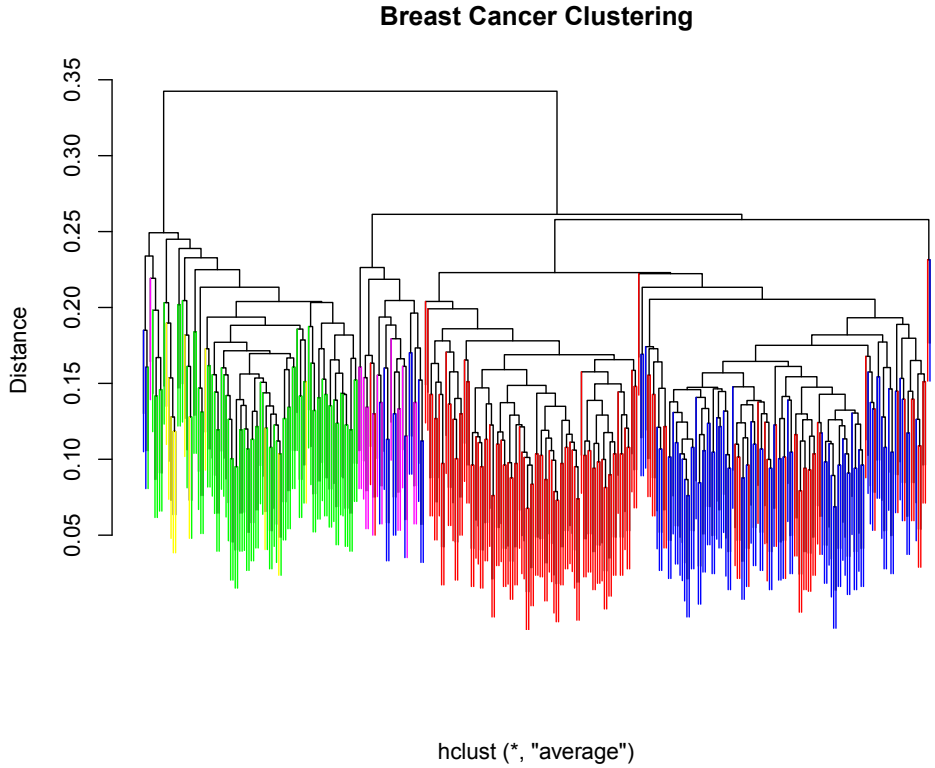

Figure 3.5: Rank-based hierarchical clustering of breast cancer tumors. Colors represent subtypes predicted by PAM50: red - Luminal A, blue - Luminal B, green - basal-like, magenta - HER2+, yellow - Normal.

subtypes of tumors are generally separated from each other where samples in the same subtype tended to have smaller distances than those in different subtypes. The distinction between luminal A and B tumors is not very clear and luminal B is the most scattered group.

To quantitatively assess concordance between PAM50 predictions and the rank-based clustering results, Cohen's Kappa (κ) was used as implemented in the R package *epibasix*. κ is a statistical measure of inter-rater agreement for categorical items and is thought to be a more robust measure than simple percent agreement because it discounts the effect of agreement by chance. κ ranges from 0 to 1 and typically has qualitative descriptions about agreement associated with different intervals. One problem of using κ is that the it assumes equal number of classes predicted by two methods for which the concordance is

assessed, and there are only four clusters identified in Figure 3.5. However, as the normal-like tumors are considered as an invalid subtype in some studies [38, 88], they have been excluded in the concordance analysis. As a result, the predictions from PAM50 and the rank-based clustering are summarized in Table 3.9.

Table 3.9: Subtype classifications by PAM50 and the rank-based clustering.

Cluster / Subtype	Lum A	Lum B	Basal	HER2+
1	94	0	0	0
2	38	89	0	0
3	0	2	80	1
4	2	12	0	15

The estimate of κ above is 0.766 (the simple percentage agreement is 0.835), which suggests excellent agreement. Basal-like tumors are the most concordant subtype identified, and the least concordant distinction is between luminal A and B tumors. These results confirm the findings of previous studies about the basal-like and luminal cancers [46]. The subdivision of luminal tumors has been reported to be strongly dependent on the expression of proliferation-related genes that actually forms a continuum, and the identification of subclasses is hence challenging.

Subtype Prediction and Survival Analysis

Hierarchical clustering can only be applied retrospectively on a collection of samples, and classification models need to be developed to produce predictions for individual samples. Motivated by the rank-based clustering method, the k-TSP classifier is introduced as a subtype predictor. In rank-based clustering, two samples from the same cluster/class tend to have a relatively small distance and they generally possess a large

number of "consistent" gene pairs (that is, two genes preserve their relative ordering across these samples). As a result, the probability of having a "consistent" gene pair turns out to be high for two samples with a small distance. Therefore, it is natural to consider differentiating one class from another in the same way as the TSP classifier.

An attempt was made to build a subtype classifier using the clustering result on VDX. For each cluster in Table 3.9, the subtypes with the most samples was chosen to constitute the training set. In addition, 11 normal-like tumors contained in cluster 3 were also selected. As a result, the training set contains 289 samples with 94 luminal A, 89 luminal B, 80 basal-like, 15 HER2+ and 11 normal-like tumors. The k-TSP classifier was constructed with $k \leq 50$, and the multiclass problem was handled by an ensemble of five decision rules. To be precise, for a certain k, a list of top k scoring pairs was obtained for distinguishing each subtype from all other groups. The aggregation of these five lists formed the final decision rule. For predicting a sample, the number of "consistent" gene pairs between the sample and each of the five lists was counted, and the predicted subtype was taken to be the list having the most counts.

To find the optimal k, the re-substitution errors of different k-TSP classifiers were assessed on the training set and compared in Figure 3.6. The misclassification rate was used for the measure of error and the re-substitution error could be a good approximate of the generalization error when sample size is large. A similar procedure was employed in [59] for the model optimization of k-TSP. The minimum error was achieved by $k = 48, 49$ and 50. The 48-TSP classifier was chosen because it contained less genes in total.

The subtype classifier was tested on an independent validation set combining 398 samples from MAINZ and TRANSBIG (Table 3.8). There are 201 luminal A, 121 luminal B, 57 basal, 11 HER2+ and 8 normal-like tumors predicted. This result is compared to

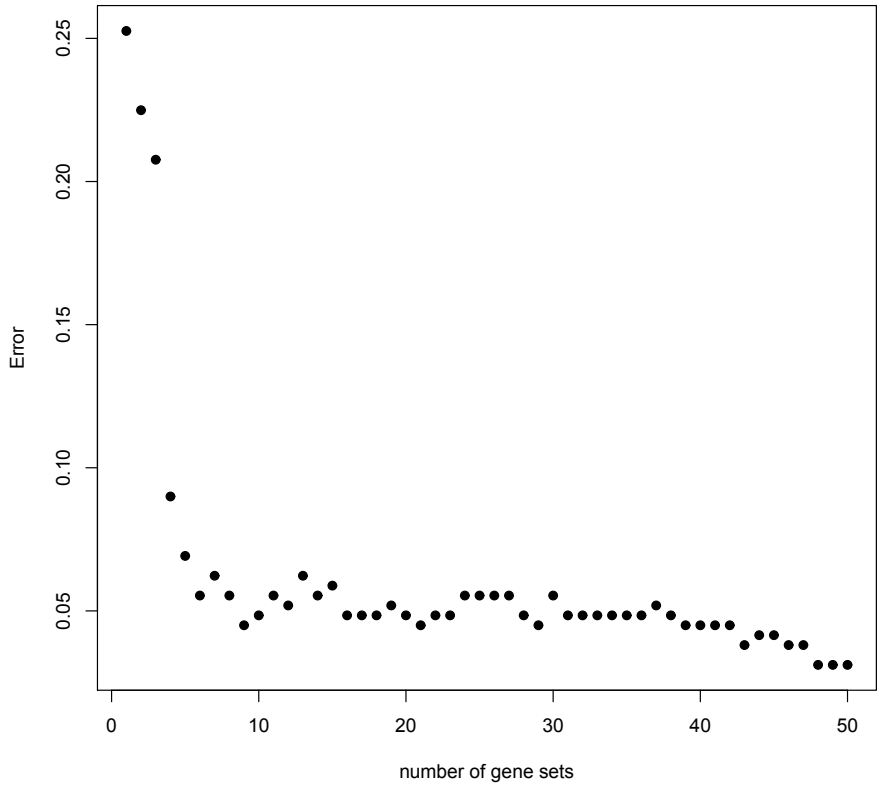

Figure 3.6: Re-substitution errors of k-TSP classifiers.

the predictions of PAM50 in Table 3.10. Excellent agreement is observed between PAM50 and 48-TSP with $\kappa = 0.779$.

Table 3.10: Subtype predictions by PAM50 and the 48-TSP classifier.

48-TSP / PAM50	Lum A	Lum B	Basal	HER2+	Normal
Lum A	183	9	0	0	9
Lum B	22	93	0	6	0
Basal	0	1	53	0	3
HER2+	1	3	0	7	0
Normal	0	0	2	0	6

To analyze and compare the prognostic value of subtype predictions, Cox propor-

tional hazard regression was considered. Subtypes were treated as categorical with no assumption made on order across subtypes. Both univariate and multivariate analysis were employed. In the multivariate model, the three available clinical-pathological variables, age, tumor size (>2cm vs. <2cm) and histological grade, were included. Distant metastasis-free survival times were used as endpoints and censored at 10 years.

Table 3.11: Univariate Cox proportional hazards models of breast cancer patients.

Variables	HR	LR	Log-rank score	P value
Age	0.997	0.08	0.46	0.4955
Size	1.37	2.64	2.74	0.098
Grade	1.38	4.74	13.01	0.0003
48-TSP		20.45	27.05	0.0004
Luminal A	1			-
Luminal B	1.89			0.0111
Basal-like	2.54			0.0014
HER2+	1.75			0.3528
Normal-like	6.71			<0.0001
PAM50		21.9	23.17	0.0002
Luminal A	1			-
Luminal B	2.51			0.0003
Basal-like	2.63			0.0014
HER2+	1.51			0.4936
Normal-like	4.07			0.0004

In the univariate analysis, the 48-TSP and PAM50 classifier are both found to be significantly associated with distant metastasis-free survival. 48-TSP gives a lower likelihood ratio but a higher log-rank test score. The luminal A versus luminal B segregation

is found more significantly in PAM50 than for 48-TSP. The basal-like segregation for both predictors have the same significance level. 48-TSP seems to have a better segregation for normal-like tumors. The luminal A versus HER2+ segregation is not found significantly by both classifiers. Besides the subtype variable, only histological grade appears to be significantly associated with distant metastasis-free survival. In the multivariate result, both predictors obtain a large increase of the likelihood ratio when adding the subtype information to the variable group of age, tumor size and histological grade.

Survival curves were plotted using the Kaplan-Meier estimator and compared using the log-rank test. The subtype distinctions in both curves are statistically significant, which confirms the substantial prognostic value of breast cancer subtyping on early untreated patients. The prognosis of patients in both curves seem to be similar and are almost consistent with risk classifications in previous studies [28]: lumina A tumors are classified as low risk while luminal B, basal-like and HER2+ tumors are classified as high risk. The abnormally high risk of the normal-like group is not observed previously and is likely to be incorrect due to the small number identified. As mentioned before, the luminal B was better separated from luminal A by PAM50, but the segregation of basal-like and HER2+ versus luminal A was better in predictions by 48-TSP.

Summary

The prognosis of breast cancer via subtype predictors has been extensively explored in the last decade. A major family of methods are the SSPs based on lists of intrinsic genes. While efforts have been made to reduce the number of intrinsic genes used from 500 to 50, the hierarchical cluster analysis employed by all SSPs has several limitations when applied to gene expression data from different cohorts and microarray technologies. More

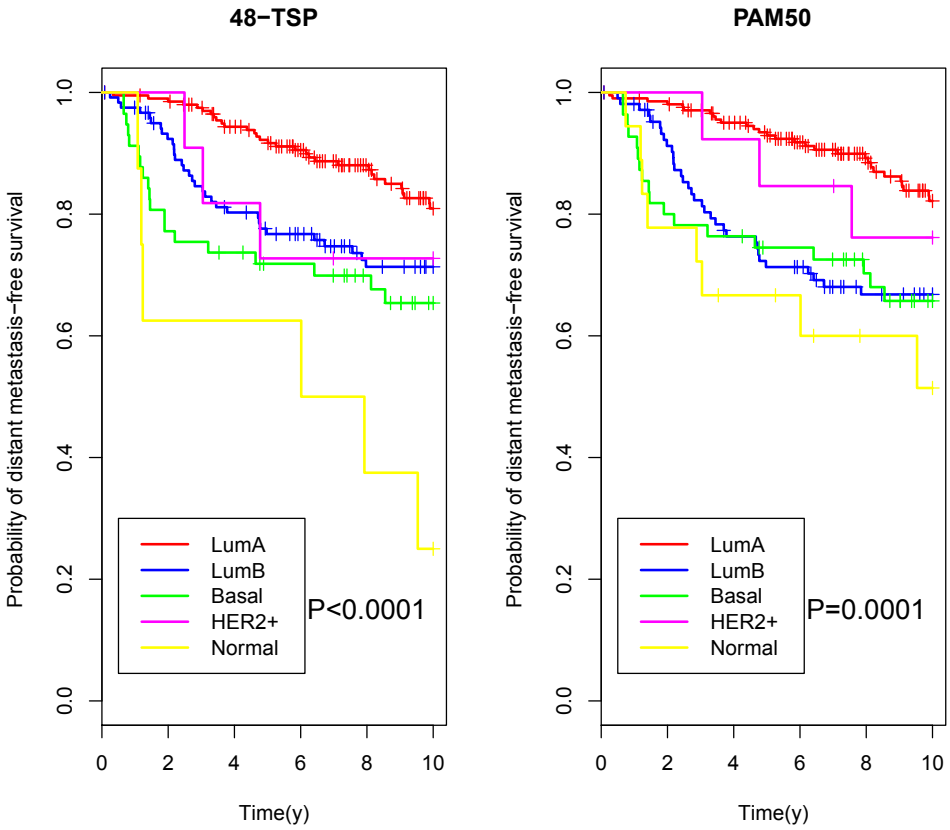

Figure 3.7: Kaplan-Meier curves.

importantly, all SSPs adopt similar procedures to identify subtypes of breast cancer, the results need to be validated by other independently developed methods. In the study of Haibe-Kains et al. [38], the SSPs only yielded fair to moderate concordance with another family of methods called Subtype Classification Models.

The rank-based clustering approach described in this chapter provides an independent validation on a collection of three gene expression data sets of breast cancer patients. This approach adopts the Kendall's rank coefficient as the distance metric in hierarchical cluster analysis, and can be robust against some commonly used preprocessing techniques required for microarray data. In this chapter, the k-TSP classifier is proposed to predict subtypes of individual samples as compared to the nearest centroid method used in SSPs.

In particular, the prognostic value of early untreated patients is investigated.

The rank-based clustering approach on 344 breast cancer patients of VDX produced four large clusters using 44 intrinsic genes. PAM50 predictions were used to identify the subtype associated with each cluster. Basal-like tumors seem to be significantly different in terms of relative orderings of genes and are all contained in one cluster. The segregation of other subtypes is less significant, especially for luminal A and B tumors. The PAM50 predictions and the rank-based clustering results have obtained excellent concordance with Cohen's Kappa equal to 0.766. 289 samples in the same cluster and also predicted by PAM50 as the same subtype were used to train a k-TSP classifier. The choice of k was optimized on the training set using the re-substitution errors. The optimal 48-TSP classifier was validated on 398 samples a combined gene expression data from MAINZ and TRANSBIG study. In the survival analysis, the subtype predictions were found significantly associated with distant metastasis-free survival. The significance level was comparable to that of PAM50. The segregation of some survival curves (e.g., HER2+ versus luminal A) was even better than that of PAM50.

This study demonstrates the ability of the rank-based clustering approach for identifying breast cancer subtypes. Samples from different subtypes tend to have different pairwise orderings, which can be a robust signal for distinguishing subtypes. The high concordance with the subtypes identified by PAM50 provides an independent validation for the existence of these intrinsic subtypes. Also, the predictive power of the k-TSP classifier for individual samples has been demonstrated through the survival analysis.

Chapter 4

Discussion and Conclusions

4.1 Multiclass Relative Expression Analysis

This thesis is focused on the development of multiclass rank-based methodologies to address the current limitations of applying statistical learning techniques on microarray gene expression data. Methods are tailored to the realities (e.g., "small n, large p") of the available data and to the properties (e.g., multiclass, biological) of problems. The relative expression orderings of small-sized gene sets are explored and modeled for microarray-based diagnosis and prognosis of human diseases.

In Chapter 2, the Top Scoring Set method is developed for classification of multiclass problems with respect to human diseases such as cancer. Specific m-gene sets named "top scoring sets" are sought to distinguish m classes based on the relative comparison of expression values. The underlying assumption is that the high expression level of a certain gene relative to others in each top scoring set is strongly indicative of the sample from a distinct class. The assumption leads to a transparent and potentially interpretable decision rule that often involves a small collection of m-gene sets. But the search process

for such gene sets turns out to be a main challenge. Univariate statistical tests (e.g., Wilcoxon rank sum, ANOVA) are commonly used for efficient dimension reduction, but they are not best suited for the TSS classifier as well as the underlying biology of the problem under study. As a result, the greedy search algorithm is proposed to pre-select a collection of locally optimal gene sets in which the top scoring sets are sought. This pre-selection process adopts an iterative procedure and enables a fast search among the large feature space. In addition, ensemble classification is also investigated for the TSS approach. Three ensemble methods (k-TSS, RF-TSS and SAMME-TSS) are introduced and studied. In particular, k-TSS considers a majority rule based on the top k scoring sets; RF-TSS uses a large collection of randomly formed gene sets and SAMME-TSS adopts the SAMME algorithm, a multiclass extension of the Adaboost algorithm. The methods proposed in this Chapter are validated on a number of gene expression data. The greedy search-based TSS method are tested on seven disease classification problems and the predictive performance is compared to that of five benchmark classifiers (kNN, NB, RF, l-SVM and PAM). Moreover, it has been validated on an extremely large data set from the MILE project and is compared to two different ensembles of support vector machine classifiers. In both cases, comparable or even better results have been observed. In addition, three ensemble approaches are validated on a large group of bladder cancer patients from multiple locations. Both k-TSS and SAMME-TSS are found to improve the performance of the TSS classifier.

In Chapter 3, the search process of the TSS classifier is further explored in the situation that publicly available pathway databases can be introduced for biologically meaningful dimension reduction. The formation of each gene set is restricted to contain only genes from the same pathway defined by a certain database such as KEGG or Gene Ontology.

The search efficiency is significantly improved due to the small sizes of pathways, and the relative expression comparison can be more biological interpretable than that of randomly formed gene sets. The k-TSS approach is proposed to construct an ensemble classifier in which the decision is based on top k scoring sets sought within pathways, and an acceleration algorithm is developed to improve the efficiency for optimizing k in the training process. The pathway-based k-TSS approach is validated on two well-studied microarray data sets. One is the group of acute leukemia samples from the Golub's study and the other is the large leukemia cohort from the MILE project. The effectiveness and robustness of the pathway-based TSS classifier is demonstrated under both small and large sample situation.

Chapter 3 also explores the ability of rank-based methods for unsupervised learning. A rank-based clustering approach is considered by using hierarchical cluster analysis and the Kendall's rank coefficient as the distance metric. The metric is proportional to the number of gene pairs that have the same relative ordering across two samples where gene pairs are selected from a pre-defined group of genes (e.g., pathway). Compared to the value-based metric such as the Pearson's correlation or the Euclidean distance, this rank-based metric can be more robust due to the use of relative ordering instead of the actual expression values. This clustering approach aims to detect samples having similar combinatorial behaviors in a group of genes, which can be a way to probe the dependency structure in genetic networks. This technique is used to discover the intrinsic subtypes of breast cancer tumors, and is compared to the subtypes predicted by the popular PAM50 model. The high concordance measured by Cohen's Kappa is observed between subtypes identified by two methods, which provides an independent validation for the existence of these intrinsic subtypes. In addition, the k-TSP classifier is introduced for predicting

subtypes of individual samples. The predictions of k-TSP and PAM50 on 398 early untreated breast cancer patients are compared and analyzed. The high concordance is again observed, and both models are found significantly associated with the times of distant metastasis-free survival of patients.

4.2 Potential Future Work

The Top Scoring Set approach is one of the first attempts of using the relative expression analysis in multiclass problems. In binary microarray-based classification, the TSP classifier has achieved considerable success using pairs of genes. The extension of this rank-based method in the multiclass setting is difficult. The complexity of multiclass classification is challenging for developing such rank-based method that can have a transparent decision rule involving only a small number of genes. However, this thesis demonstrates the predictive power existing in small-sized gene sets through relative expression comparison. The empirical distribution of such comparison seems to be one of statistics that can be robustly estimated from the data available. Also, the exploration of relative expression comparison is a practical attempt to study genetic interactions in biological networks, which are well-recognized to be involved in complex diseases.

Considering gene sets with the "maximum" gene changing over classes is one of many ways to investigate possible perturbations in genetic networks through expression relative comparison. An immediate extension is to consider the same property with the "minimum" gene for distinguishing classes. However, the "minimum" gene is often the gene not being expressed and is not necessarily associated with the phenotypes under study, which may not be used as a signal for classification. Although the work is not shown in this thesis, an attempt has been made to search top scoring sets in terms of the "min-

imum" gene and compare their predictive ability with those based on the "maximum" gene. The results seem to be more favorable for the "max" version of the TSS classifier. In addition, the number of complete orderings (permutations) of even a few genes is large (e.g., 4!=24, 5!=120), and it is often impractical to estimate the distribution of the complete orderings. TSS provides a way to combine some of these orderings to gain statistical significance. As the sample size grows, more orderings can be considered in the modeling process, which would give a more accurate estimate of the statistical dependency structures among genes.

The investigation of the search process for top scoring sets is another focus in this thesis. The greedy search algorithm provides a feasible way to efficiently search the entire feature space for candidates of top scoring sets. But it is generally not efficient enough on large data sets. The pathway-based search can be an excellent option for dimension reduction. The KEGG database is used throughout this thesis to define gene groups, but the use of other databases remains unexplored. One good alternative is to use the Gene Ontology database that contains far more gene groups than KEGG. For specific problems, some manually curated pathways can even be more valuable. Also, a possible future goal is to develop a rank-based enrichment analysis that compares different pathways according to their abilities of separating classes using the relative expression analysis. In this situation, the pathways with genetic perturbations that cause a certain disease are expected to be significantly associated with that disease.

The exploration of the rank-based clustering approach in Chapter 3 is far from complete. The method has the limitation that a cluster detected in the dendrogram can not be directly assigned to a well-defined subtype of breast cancer in previous studies. In fact, as discussed in [58], this limitation of class discovery exists for all methods that use hier-

archical clustering, and it can cause disagreement of subtype identification. In this thesis, only samples identified by both methods (the rank-based clustering and PAM50) as the same subtype are used for training the subtype classifier. A potential future research is to discover breast cancer subtypes directly from the clustering results with the aid of biological knowledge such as known genes or pathways associated with each subtype. Also, the distance matrix used in the clustering process is based on the list of intrinsic genes in PAM50. A possibly more objective approach would be to construct the distance matrix using genes that are known to be involved in the progression of breast cancer tumors. The subtype discovery process needs to be similar to that employed in [62, 77] in order to validate the rank-based clustering results. Once the subtypes are identified, the k-TSP classifier can be trained and tested, and compared to popular subtype models such as SSPs and SCMGENE.

At a higher level, the possible biological interpretation of relative expression comparison still needs to be validated, and the link between such comparison and the underlying mechanism causing different phenotypes is not clear. A recent study [47] has shown the effectiveness of such relative comparison based on expression levels of peptides and the associated proteins for disease diagnosis. More future studies are expected for validating the use of gene expression microarray data, and for exploring the application of the relative expression analysis to other types of data.

Bibliography

[1] Bahman Afsari. *Modeling cancer phenotypes with order statistics of transcript data*. PhD thesis, The Johns Hopkins Unversity, 2013.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, 2000.

[3] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–7, 2002.

[4] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–9, 2005.

[5] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek,

L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8):816–24, 2002.

[6] L. Breiman. *Bias, variance and arcing classifiers*. Technical report, Statistics Department, University of California, Berkerly, 1996.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] D. J. Burgess. Cancer genetics: Initially complex, always heterogeneous. *Nature Reviews Cancer*, 11:153, 2011.

[9] M. H. Cheok, W. Yang, C. H. Pui, J. R. Downing, C. Cheng, C. W. Naeve, M. V. Relling, and W. E. Evans. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nature Genetics*, 34(1):85–90, 2003.

[10] C. Chih-Chung and L. Chih-Jen. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3)(27), 2011.

[11] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[12] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[13] E. Dehan, A. Ben-Dor, W. Liao, D. Lipson, H. Frimer, S. Rienstein, D. Simansky, M. Krupsky, P. Yaron, E. Friedman, G. Rechavi, M. Perlman, A. Aviram-Goldring, S. Izraeli, M. Bittner, Z. Yakhini, and N. Kaminski. Chromosomal aberrations and

gene expression profiles in non-small cell lung cancer. *Lung Cancer*, 56(2):157–84, 2007.

[14] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14(4):457–60, 1996.

[15] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, 14(16):5158–65, 2008.

[16] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, and C. Sotiriou. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13(11):3207–14, 2007.

[17] M. Dettling and P. Buhlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–9, 2002.

[18] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[19] T. G. Dietterich and E. B. Kong. *machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Technical report, Department of Computer Science, Oregon State University, 1995.

[20] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin. Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.

[21] L. Dyrskjot, T. Thykjaer, M. Kruhoffer, J. L. Jensen, N. Marcussen, D. S. Hamilton, H. Wolf, and T. F. Orntoft. Identifying distinct classes of bladder carcinoma using microarrays. *Nature Genetics*, 33(1):90–6, 2003.

[22] L. Dyrskjot, K. Zieger, F. X. Real, N. Malats, A. Carrato, C. Hurst, S. Kotwal, M. Knowles, P. U. Malmstrom, M. de la Torre, K. Wester, Y. Allory, D. Vordos, A. Caillault, F. Radvanyi, A. M. Hein, J. L. Jensen, K. M. Jensen, N. Marcussen, and T. F. Orntoft. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. *Clinical Cancer Research*, 13(12):3545–51, 2007.

[23] J. A. Eddy, J. Sung, D. Geman, and N. D. Price. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technology in Cancer Research and Treatment*, 9(2):149–59, 2010.

[24] L. B. Edelman, G. Toia, D. Geman, W. Zhang, and N. D. Price. Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics*, 10:583, 2009.

[25] B. Efron and R. Tibshirani. On testing the significance of sets of genes. Technical report, Stanford University, August 2006. http://www-stat.stanford.edu/∼tibs/GSA/.

[26] S. Efroni, C. F. Schaefer, and K. H. Buetow. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, 2(5):e425, 2005.

[27] D. J. Erle and Y. H. Yang. Asthma investigators begin to reap the fruits of genomics. *Genome Biology*, 4(11):232, 2003.

[28] C. Fan, D. S. Oh, L. Wessels, B. Weigelt, D. S. Nuyten, A. B. Nobel, L. J. van't Veer, and C. M. Perou. Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine*, 355(6):560–9, 2006.

[29] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[30] J. Friedman. *Another approach to polychotomous classification.* Technical report, Department of Statistics, Stanford, Palo Alto, CA, 1996.

[31] L. M. Fu and C. S. Fu-Liu. Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. *FEBS Lett*, 561(1-3):186–90, 2004.

[32] M. L. Gatza, J. E. Lucas, W. T. Barry, J. W. Kim, Q. Wang, M. D. Crawford, M. B. Datto, M. Kelley, B. Mathey-Prevot, A. Potti, and J. R. Nevins. A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 107(15):6994–9, 2010.

[33] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow. Classifying gene expression profiles from pairwise mrna comparisons. *Statistical Applications in Genetics and Molecular Biology*, 3(19), 2004.

[34] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[35] T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, 1999.

[36] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005.

[37] T. Haferlach, A. Kohlmann, L. Wieczorek, G. Basso, G. T. Kronnie, M. C. Bene, J. De Vos, J. M. Hernmandez, W. K. Hofmann, K. I. Mills, A. Gilkes, S. Chiaretti, S. A. Shurtleff, T. J. Kipps, L. Z. Rassenti, A. E. Yeoh, P. R. Papenhausen, W. M. Liu, P. M. Williams, and R. Foa. Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. *Journal of Clinical Oncology*, 28(15):2529–37, 2010.

[38] B. Haibe-Kains, C. Desmedt, S. Loi, A. C. Culhane, G. Bontempi, J. Quackenbush, and C. Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104(4):311–25, 2012.

[39] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–60, 1950.

[40] D. Hand and R. J. Till. A simple generalization of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[41] M.A. et al. Harris. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32:D258–61, 2004.

[42] T. Hastie, R. Tibshirani, and J. Friedman. *The element of statistical learning*. Springer Series in Statistics, New York, NY, 2001.

[43] C. W. Hsu and C. J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:1045–1052, 2002.

[44] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

[45] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Research*, 32:D277–80, 2004.

[46] A. V. Kapp, S. S. Jeffrey, A. Langerod, A. L. Borresen-Dale, W. Han, D. Y. Noh, I. R. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7:231, 2006.

[47] P. Kaur, D. Schlatzer, K. Cooke, and M. R. Chance. Pairwise protein expression classifier for candidate biomarker discovery for early detection of human disease prognosis. *BMC Bioinformatics*, 13:191, 2012.

[48] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[49] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–9, 2001.

[50] S. Kim, M. Kon, and C. DeLisi. A pathway-based classification of human breast cancer. *Biology Direct*, 7:21, 2012.

[51] R. Kohavi and D. H. Wolpert. Bias plus variance decomposition for zero-one loss functions. *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283, 1996.

[52] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217, 2008.

[53] J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

[54] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–9, 2010.

[55] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–37, 2004.

[56] X. Lin, B. Afsari, L. Marchionni, L. Cope, G. Parmigiani, D. Naiman, and D. Geman. The ordering of expression among a few genes can provide simple cancer biomarkers and signal brca1 mutations. *BMC Bioinformatics*, 10:265, 2009.

[57] X. Lin, B. Afsari, L. Marchionni, L. Cope, G. Parmigiani, D. Q. Naiman, and D. Geman. The ordering of expression among a few genes can provide simple cancer biomarkers and signal brca1 mutations. *BMC Bioinformatics*, 10:256, 2009.

[58] A. Mackay, B. Weigelt, A. Grigoriadis, B. Kreike, R. Natrajan, R. A'Hern, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. *Journal of the National Cancer Institute*, 103(8):662–73, 2011.

[59] L. Marchionni, B. Afsari, D. Geman, and J. T. Leek. A simple and reproducible breast cancer prognostic test. *BMC Genomics*, 14:336, 2013.

[60] M. Mramor, G Leban, J Demsar, and B Zupan. Visualization-based cancer microarray data classification analysis. *Bioinformatics*, 23(16):2147–54, 2007.

[61] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*, 351(27):2817–26, 2004.

[62] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–7, 2009.

[63] S. K. Patnaik, E. Kannisto, S. Knudsen, and S. Yendamuri. Evaluation of microrna expression profiles that may predict recurrence of localized stage i non-small cell lung cancer after surgical resection. *Cancer Research*, 70(1):36–45, 2010.

[64] C. M. Perou, J. S. Parker, A. Prat, M. J. Ellis, and P. S. Bernard. Clinical implemen-

tation of the intrinsic subtypes of breast cancer. *The Lancet Oncology*, 11(8):718–9, 2010.

[65] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, A. L. Lonning, P. E. amd Borresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–52, 2000.

[66] M. Pirooznia, Y. J. Yang, and M. Q. Yang. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(Suppl 1), 2008.

[67] L. Pusztai, C. Mazouni, K. Anderson, Y. Wu, and W. F. Symmans. Molecular classification of breast cancer: limitations and potential. *The Oncologist*, 11(8):868–77, 2006.

[68] J. Quackenbush. Microarray analysis and tumor classification. *New England Journal of Medicine.*, 354(23):2463–72, 2006.

[69] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[70] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[71] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and

J. Quackenbush. Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374–8, 2003.

[72] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70, 1995.

[73] M. Schmidt, D. Bohm, C. von Torne, E. Steiner, A. Puhl, H. Pilch, H. A. Lehr, J. G. Hengstler, H. Kolbl, and M. Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13):5405–13, 2008.

[74] B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge, MA, 2002.

[75] M. A. Shah, R. Khanin, L. Tang, Y. Y. Janjigian, D. S. Klimstra, H. Gerdes, and D. P. Kelsen. Molecular classification of gastric cancer: a new paradigm. *Clinical Cancer Research*, 17(9):2693–701, 2011.

[76] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–74, 2001.

[77] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O.

Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8418–23, 2003.

[78] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Journal of the National Cancer Institute*, 98(4):262–72, 2006.

[79] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–43, 2005.

[80] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9:319, 2008.

[81] J. Su, B. J. Yoon, and E. R. Dougherty. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*, 4(12):e8161, 2009.

[82] A. Subramanian, Tamayo. P., V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

[83] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman. Simple decision

rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–904, 2005.

[84] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–9, 2005.

[85] R. Tibshirani, T. Hastie, B Narasimhan, and G Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6567–72, 2002.

[86] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend.

[87] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–9, 2005.

[88] B. Weigelt, A. Mackay, R. A'hern, R. Natrajan, D. S. Tan, M. Dowsett, A. Ashworth, and J. S. Reis-Filho. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology*, 11(4):339–49, 2010.

[89] L. Xu, D. Geman, and Winslow R. L. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, 8:275, 2007.

[90] Lin Xue. *Rank-based methods for statistical analysis of gene expression microarray data*. PhD thesis, The Johns Hopkins Unversity, 2008.

[91] S. Yang and D. Q. Naiman. Multiclass cancer classification based on gene expression comparison. *Statistical Applications in Genetics and Molecular Biology*, 13(4):477–96, 2014.

[92] L. Ying and H. Jiawei. Cancer classification using gene expression data. *Information Systems*, 28:243–268, 2003.

[93] J.M. Zahn, S. Poosala, A. B. Owen, D. K. Ingram, A. Lustig, A. Carter, A. T. Weeraratna, D. D. Taub, M. Gorospe, K. Mazan-Mamczarz, E. G. Lakatta, K. R. Boheler, X. Xu, M. P. Mattson, G. Falco, M. S. Ko, D. Schlessinger, J. Firman, S. K. Kummerfeld, W. H. Wood, A. B. Zonderman, S. K. Kim, and K. G. Becker. Agemap: a gene expression database for aging in mice. *PLoS Genetics*, 3(11):e201, 2007.

[94] H. Zhao, C. J. Logothetis, and I. P. Gorlov. Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. *Prostate Cancer and Prostatic Diseases*, 13(3):252–9, 2010.

[95] J. Zhu, H. Zou, S. Rosset, and T. Hastie. Multi-class adaboost. *Statistics and Its Inferences*, 2:349–360, 2009.

Vita

Sitan Yang was born on October 11, 1985 in Chongqing, China. He graduated from Yucai middle school and attended Sichuan University in Chengdu since 2004. He obtained his B.S. in Mathematics and Applied Mathematics in 2008. At that time, he received the excellent dissertation award and was the recipient of the outstanding student fellowship for three consecutive years. Also, he was the meritorious winner in the 2007 Mathematical Contest in Modeling held in USA. Sitan began his graduate study in the Department of Applied Mathematics and Statistics at Johns Hopkins University in 2008 and earned his M.S.E in 2010. His research focuses on developing new statistical learning methods for microarray data analysis.